frontiers | Research Topics

Vetinformatics: An insight for decoding livestock systems through in silico biology

Edited by
Jun-Mo Kim and Rajesh Kumar Pathak

Published in
Frontiers in Veterinary Science

ISSN 1664-8714
ISBN 978-2-8325-3913-2
DOI 10.3389/978-2-8325-3913-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Vetinformatics: An insight for decoding livestock systems through in silico biology

Topic editors

Jun-Mo Kim — Chung-Ang University, Republic of Korea
Rajesh Kumar Pathak — Chung-Ang University, Republic of Korea

Citation

Kim, J.-M., Pathak, R. K., eds. (2023). *Vetinformatics: An insight for decoding livestock systems through in silico biology*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3913-2

Table of contents

frontiers | Frontiers in Veterinary Science

TYPE Editorial
PUBLISHED 30 October 2023
DOI 10.3389/fvets.2023.1292733

OPEN ACCESS

EDITED AND REVIEWED BY
Mehar S. Khatkar,
The University of Sydney, Australia

*CORRESPONDENCE
Jun-Mo Kim
✉ junmokim@cau.ac.kr

RECEIVED 12 September 2023
ACCEPTED 17 October 2023
PUBLISHED 30 October 2023

CITATION
Kim J-M and Pathak RK (2023) Editorial: Vetinformatics: an insight for decoding livestock systems through *in silico* biology. *Front. Vet. Sci.* 10:1292733. doi: 10.3389/fvets.2023.1292733

Editorial: Vetinformatics: an insight for decoding livestock systems through *in silico* biology

Jun-Mo Kim* and Rajesh Kumar Pathak

Department of Animal Science and Technology, Chung-Ang University, Anseong-si, Republic of Korea

KEYWORDS

***in silico* analysis, molecular docking, molecular dynamics simulation, multiomics, network analysis, next-generation sequencing, vetinformatics**

Editorial on the Research Topic
Vetinformatics: an insight for decoding livestock systems through *in silico* biology

Computers have become an integral part of our daily lives, and we are dependent on them for many things. For instance, research is almost impossible without computers. In the early 1970s, Paulien Hogeweg and Ben Hesper coined the term "bioinformatics" (1), which became an independent discipline after its significant role in the Human Genome Project (HGP) (2). Decoding problems arising in the field of biological sciences via computation is known as bioinformatics (3).

With the world's population booming and natural resources dwindling due to human activity, veterinary science is becoming key in research and development to meet growing demands (4, 5). To meet the growing demand, veterinary science must integrate informatics to manage complex data and enhance research and development activities in various areas of veterinary sciences (4).

To enable novel discoveries, extensive use of *in silico* tools such as BLAST (6) and databases i.e. Bovine Genome Database, Porcine Translational Research Database, *etc* are required (6–8). Moreover, computers and information science have been integrated into all aspects of the veterinary profession, leading to the concept of vetinformatics. While bioinformatics is a broad field that focuses on several areas of biology, vetinformatics is specifically concerned with addressing problems in the field of veterinary science (4).

The establishment of the Association for Veterinary Informatics took place in 1981 (https://avinformatics.org/), where it was connected veterinary with informatics, but its focus is primarily on veterinary medicine. Some articles have also been published on "veterinary informatics", where informatics has been used to advance the field of veterinary medicine (9, 10).

In 2016, Sujatha et al., published a brief review entitled "*Vetinformatics: A New Paradigm for Quality Veterinary Services*". This review attempted to highlight applications of vetinformatics in veterinary science (11). Subsequently, in 2022 we authored a comprehensive review on vetinformatics, entitled "*Vetinformatics from functional genomics to drug discovery: Insights into decoding complex molecular mechanisms of livestock systems in veterinary science*" in which we covered many aspects (4).

Accordingly, vetinformatics should also be considered as an important subject, similar to Pharmacoinformatics, Chemoinformatics, Genomeinformatics, Agriinformatics, Cropinformatics, Biomedical informatics and other informatics fields are considered. Therefore, vetinformatics is new and does not have an interesting history yet, but its approaches are important for problem solving not only in veterinary medicine but also in various areas of veterinary science (4).

In veterinary science, animal production is a highly intricate process with three basic interconnected components: animal biology, environment, and management techniques. Therefore, *in silico* approaches are required to bridge the gaps between genotype and phenotype to enhance efficiency in livestock productivity and sustainability. With the advent of several omics platforms and next-generation sequencing technologies, an enormous amount of animal data has been generated. While major bioinformatics databases and tools are available for the management and analysis of these data, veterinarians require animal and species-specific databases for effective management and future use. In addition, animal-specific tools for data analysis and integration, as well as computational and mathematical models for predicting the behavior of animal systems in different conditions, are necessary. Hence, vetinformatics has become an essential subject in the discipline of veterinary sciences, as it enables the handling and evaluation of large amounts of data and mining of important information that can aid researchers in decoding livestock systems to accelerate research and development.

Vetinformatics-related projects focus on the design and development of databases for the documentation of useful information about medicinal plants available in the literature and public domain for use in the discovery of herbal veterinary medicine. Animal genetic resource information is also being documented in the form of a database, and useful organism/species-specific databases are being created for the management of omics data sets. There is an effort to update the content available in veterinary databases for better functionality and create better and faster GUI-based data integration and analysis tools. Additionally, publicly available software is being improved for ease of use by veterinary biotechnologists and non-computer scientists and veterinarians. There is also a focus on the design and development of platform-independent software for vetinformatics research as well as vetinformatics training of undergraduate and graduate students and faculty in veterinary and animal science for analysis of multi-omics data.

The current Research Topic is "*Vetinformatics: An Insight for Decoding Livestock Systems Through In Silico Biology*". Nine out of 17 articles were accepted for publication in this special Research Topic.

The first article in this Research Topic outlines the analysis of the structural and functional properties of *Mycoplasma gallisepticum* variable lipoprotein hemagglutin (vlhA) proteins, which are crucial for immune evasion. The results suggest diverse mechanisms for vlhA protein function in immune evasion, and the predicted 3D structure can aid in understanding its interaction with other molecules (Mugunthan and Harish). The second article reports the impact of preweaning vaccination on gene expression in calves. Results show that regardless of vaccination status, there was an increase in gene expression related to specialized proresolving mediator production, lipid metabolism, and stimulation of immunoregulatory T cells, while vaccination was associated with gene expression related to natural killer cell activity and helper T-cell differentiation (Scott et al.). The third article discusses the challenges facing the livestock industry due to climate change and increased demand for food and how new scientific and technological advancements can help. It highlights the importance of vetinformatics and its potential for improving veterinary research, breeding, disease prevention, management, and sustainability (Pathak and Kim). The fourth article outlines attempts to develop a multiepitope-based vaccine candidate using major and minor capsid proteins of infectious bursal disease virus. The proposed vaccine candidate has been evaluated as antigenic, immunogenic, and non-allergenic with potential to overcome the safety and protection issues of existing live-attenuated vaccines. Further experimental studies are required to assess the efficacy of the proposed vaccine candidate *in vivo* (Gul et al.). The fifth article suggests that the uncoupling proteins (UCPs) can be functional markers for identifying metabolic state, thermogenesis, and oxidative stress in birds, and their corresponding genes could be considered as candidates for use in breeding programs aimed at balancing energy expenditure and reactive oxygen species production (Davoodi et al.). The sixth article in this Research Topic aims to evaluate the quality of reference genomes and gene annotations in 114 species. The proposed next-generation sequencing (NGS) applicability index, which integrates 10 effective indicators, can help determine technological boundaries and examine the direction of future development in each species (Park et al.). The seventh article demonstrates the consistency and variability of data produced by reference-free *de novo* transcriptomes and reference-based datasets for identifying, annotating, and analyzing genes related to four major traits of water buffalo. The findings suggest that the characterized genes will enrich the knowledge of genetics for use in molecular breeding to improve the productivity of water buffalo (Mishra et al.). Articles eight and nine in this special Research Topic highlight the importance and application of machine-learning in veterinary science. They focus on detection of malignancies in canine subcutaneous and cutaneous masses (Dank et al.) and demonstrate the automated monitoring of diseased chickens (Bakar et al.).

Considering the current scenario and the increasing demand for *in silico* tools and databases for use in veterinary science, this series presents the achievements of vetinformatics and how it will be helpful in decoding livestock systems. This is a timely and exciting opportunity to harness the potential of vetinformatics for animal health and welfare.

Author contributions

J-MK: Funding acquisition, Resources, Writing—original draft, Writing—review & editing. RKP: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C1005830).

Acknowledgments

We extend our deepest appreciation to the authors of the articles for their invaluable contributions. We also extend our gratitude to the reviewers, whose generous commitment in evaluating and providing valuable feedback has significantly contributed to the improvement of these articles. Furthermore, we are grateful to the editorial team of *Frontiers in Veterinary Science* for their assistance during the editing of this special Research Topic. The Chung-Ang University, Anseong campus is gratefully acknowledged for providing the necessary facilities.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol.* (2011) 7:e1002021. doi: 10.1371/journal.pcbi.1002021

2. Mount DW. *Bioinformatics: Sequence and Genome Analysis (2nd Ed.).* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press (2014).

3. Pathak RK, Singh DB, Singh R. Introduction to basics of bioinformatics. In: *Bioinformatics.* London: Elsevier (2022). p. 1–15. doi: 10.1016/B978-0-323-89775-4.00006-7

4. Pathak RK, Kim JM. Vetinformatics from functional genomics to drug discovery: insights into decoding complex molecular mechanisms of livestock systems in veterinary science. *Front Vet Sci.* (2022) 9:1008728. doi: 10.3389/fvets.2022.1008728

5. Alders R, de Bruyn J, Wingett K, Wong J. One health, veterinarians and the nexus between disease and food security. *Aust Vet J.* (2017) 95:451–3. doi: 10.1111/avj.12645

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* (1990) 215:403–10. doi: 10.1016/S0022-2836(05)80360-2

7. Shamimuzzaman M, Le Tourneau JJ, Unni DR, Diesh CM, Triant DA, Walsh AT, et al. Bovine genome database: new annotation tools for a new reference genome. *Nucleic Acids Res.* (2020) 48:D676–D81. doi: 10.1093/nar/gkz944

8. Dawson HD, Chen C, Gaynor B, Shao J, Urban JF. The porcine translational research database: a manually curated, genomics and proteomics-based research resource. *BMC Genomics.* (2017) 18:643. doi: 10.1186/s12864-017-4009-7

9. Bellamy J. Veterinary informatics–why are we dragging our feet? *Canad Vet J.* (1999) 40:861.

10. Lustgarten JL, Zehnder A, Shipman W, Gancher E, Webb TL. Veterinary informatics: forging the future between veterinary medicine, human medicine, and one health initiatives—a joint paper by the association for veterinary informatics (avi) and the CTSA one health alliance (Coha). *JAMIA open.* (2020) 3:306–17. doi: 10.1093/jamiaopen/ooaa005

11. Sujatha P, Kumarasamy P, Preetha S, Balachandran P. Vetinformatics: a new paradigm for quality veterinary services. *Res Rev J Vet Sci Technol.* (2016) 5:16–9. doi: 10.37591/rrjovst.v5i2.537

TYPE Original Research
PUBLISHED 04 August 2022
DOI 10.3389/fvets.2022.943831

Check for updates

OPEN ACCESS

EDITED BY
Rajesh Kumar Pathak,
Chung-Ang University, South Korea

REVIEWED BY
Dev Bukhsh Singh,
Siddharth University Kapilvastu, India
Neetesh Pandey,
Indraprastha Institute of Information Technology Delhi, India
Pallavi Shah,
Indian Veterinary Research Institute (IVRI), India

*CORRESPONDENCE
Mani Chandra Harish
mc.harishin@tvu.edu.in

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

RECEIVED 14 May 2022
ACCEPTED 01 July 2022
PUBLISHED 04 August 2022

CITATION
Mugunthan SP and Harish MC (2022) *In silico* structural homology modeling and functional characterization of *Mycoplasma gallisepticum* variable lipoprotein hemagglutin proteins. *Front. Vet. Sci.* 9:943831. doi: 10.3389/fvets.2022.943831

In silico structural homology modeling and functional characterization of *Mycoplasma gallisepticum* variable lipoprotein hemagglutin proteins

Susithra Priyadarshni Mugunthan and Mani Chandra Harish*

Department of Biotechnology, Thiruvalluvar University, Vellore, India

Mycoplasma gallisepticum variable lipoprotein hemagglutin (vlhA) proteins are crucial for immune evasion from the host cells, permitting the persistence and survival of the pathogen. However, the exact molecular mechanism behind the immune evasion function is still not clear. *In silico* physiochemical analysis, domain analysis, subcellular localization, and homology modeling studies have been carried out to predict the structural and functional properties of these proteins. The outcomes of this study provide significant preliminary data for understanding the immune evasion by vlhA proteins. In this study, we have reported the primary, secondary, and tertiary structural characteristics and subcellular localization, presence of the transmembrane helix and signal peptide, and functional characteristics of vlhA proteins from *M. gallisepticum* strain R low. The results show variation between the structural and functional components of the proteins, signifying the role and diverse molecular mechanisms in functioning of vlhA proteins in host immune evasion. Moreover the 3D structure predicted in this study will pave a way for understanding vlhA protein function and its interaction with other molecules to undergo immune evasion. This study forms the basis for future experimental studies improving our understanding in the molecular mechanisms used by vlhA proteins.

KEYWORDS

variable lipoprotein hemagglutin, immune evasion, bioinformatics, avian mycoplasmosis, *M. gallisepticum*

Introduction

The bacteria of class Mollicutes are described as simplest self–replicating life forms due to their small cell size and complete lack of cell wall, limited metabolic pathway and reduced genome size (1). The *Mycoplasmataceae* family in Mollicutes includes majority of disease causing pathogens in medical and veterinary fields. A great number of *Mycoplasma* species are pathogenic to humans and animals which cause chronic infections consequential in infectious diseases. To adapt and survive the challenging and complex host environment, the mycoplasmas use combinational genetic machinery for phase and size variation of major surface components. Due to the lack of cell wall, the outer surface of the mycoplasma membrane plays a crucial role in the infection process, transport of nutrients, interaction with host cells, and host immune defense. Thus, gaining knowledge in the process of how and when the antigenic variation occurs can offer important insights to the tactics used by mycoplasmas to cause infection in host cells.

Mycoplasma gallisepticum is one of the most important avian pathogens which causes chronic respiratory disease (CRD) in chickens with the symptoms of cough, nasal discharge, low appetite, reduced hatchability and chick viability, loss of weight, and decreased egg production (1, 2). The responsible pathogenic events are due to genes that encode cytoadhesion and surface components with antigenic variation which involves the immune evasion of the host (3). *M. gallisepticum* infection results in infectious sinusitis in turkeys (swollen infraorbital sinuses) and conjunctivitis in finches.

The immune evasion of *M. gallisepticum* is regulated by the vlhA gene family. This family consists of 43 vlhA genes located in five loci (Table 1). The major function of this gene family is to engender antigenic diversity which assists in immune evasion during infection. The vlhA gene family shows phase variation during acute phase and immune evasion during the chronic phase of infection (4, 5). The phase variation may occur impulsively or by an immune attack and is crucial for survival of *M. gallisepticum* in host cells (6–8). Various mechanisms for phase variation like gene conversion, site specific recombination, DNA slippage, and reciprocal recombination were utilized by different species of Mycoplasma (9). The vlhA gene products are speculated to be engaged in the attachment of host apolipoprotein A1 (10, 11) and red blood cells (12). The phase variation of *M. gallisepticum* is exclusive and has not been studied yet. Among the other vlhA genes, vlhA 3.03, 2.02 and 4.01 genes are primarily expressed in the initial phase of infection, whereas vlhA 1.07 and 5.13 are expressed in the later stages of infection. The prototype followed by *M. gallisepticum* to express the dominant vlhA gene during the course of infection is stochastic and the mechanism is unknown and yet to be explored (4). This study employed computational tools to understand the evolutionary relationship of the vlhA proteins; structural studies which include its primary sequence analysis, and secondary and tertiary structural analysis, functional studies like the cellular localization, presence of the transmembrane helix and signal peptide in vlhA proteins, and finally identification of functional domain were performed. To date, no *in silico* structural and functional studies have been reported for *M. gallisepticum* vlhA proteins. The diagrammatic representation of the work flow is presented in Figure 1. The list of bioinformatics tools and servers employed in this study is given in Table 2.

TABLE 1 List of vlhA genes based on their group analyzed in this study.

vlhA 1	vlhA 2	vlhA 3	vlhA 4	vlhA 5
vlhA.1.01	vlhA.2.01	vlhA.3.01	vlhA.4.01	vlhA.5.01a
vlhA.1.02	vlhA.2.02	vlhA.3.02	vlhA.4.02	vlhA.5.01b
vlhA.1.03		vlhA.3.03	vlhA.4.03	vlhA.5.01c
vlhA.1.04		vlhA.3.04	vlhA.4.04	vlhA.5.02
vlhA.1.05		vlhA.3.05	vlhA.4.05	vlhA.5.03
vlhA.1.06		vlhA.3.06	vlhA.4.06	vlhA.5.04
vlhA.1.07		vlhA.3.07	vlhA.4.07	vlhA.5.05
vlhA.1.08		vlhA.3.08	vlhA.4.07.1	vlhA.5.06
vlhA.1.08b		vlhA.3.09	vlhA.4.07.2	vlhA.5.07
			vlhA.4.07.4	vlhA.5.08
			vlhA.4.07.6	vlhA.5.09
			vlhA.4.08	vlhA.5.10a
			vlhA.4.09	vlhA.5.10b
			vlhA.4.10	vlhA.5.11
			vlhA.4.11	vlhA.5.12
			vlhA.4.12	vlhA.5.13

Understanding the structural and functional properties of vlhA proteins of *M. gallisepticum* will provide the first step/lead in the direction of understanding of underlying molecular mechanisms involved. In this study, we used *in silico* methods to determine the physical, structural, and functional characteristics of vlhA proteins.

Materials and methods

Sequence retrieval

The amino acid sequences of vlhA proteins from *Mycoplasma gallisepticum* strain R low used in this study were retrieved from UniProt in the FASTA format. The protein names and their unique UniProt IDs are shown in Supplementary Table 1.

Phylogenetic analysis

To understand the evolutionary relationships between the vlhA proteins, a phylogenetic tree was constructed using Phylogeny.fr, online software for phylogenetic analysis (13). The "One Click" option was used where the alignment was performed by MUSCLE, curation was performed by Gblocks, phylogeny was performed by PhyML, and Tree Rendering was performed by TreeDyn.

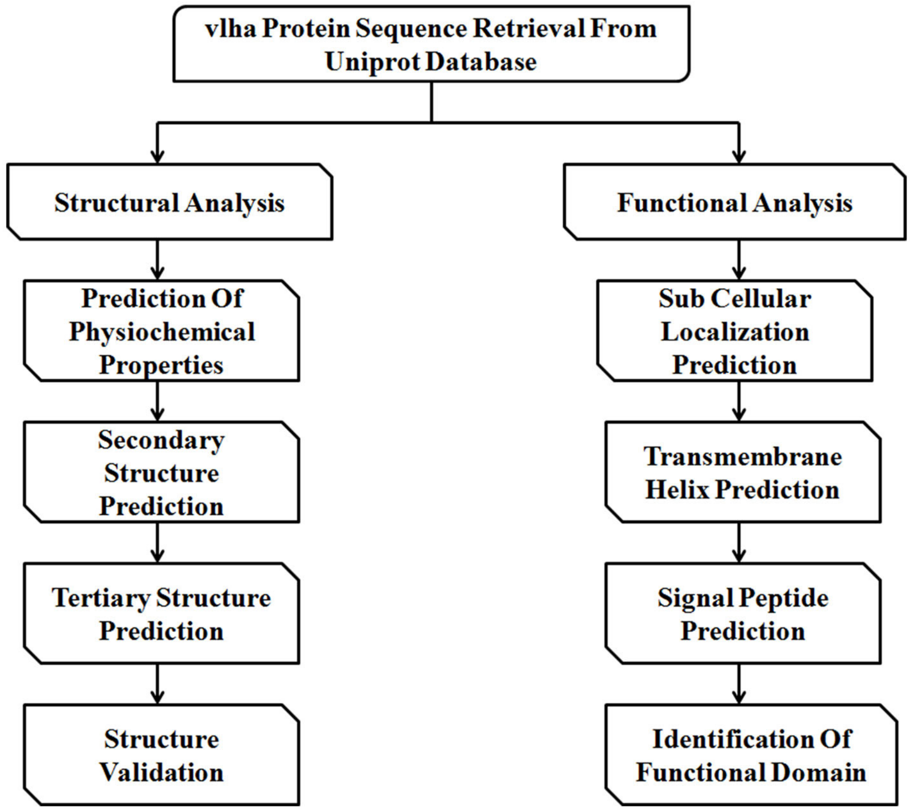

vlha Protein Sequence Retrieval From Uniprot Database
Structural Analysis
Prediction Of Physiochemical Properties
Secondary Structure Prediction
Tertiary Structure Prediction
Structure Validation
Functional Analysis
Sub Cellular Localization Prediction
Transmembrane Helix Prediction
Signal Peptide Prediction
Identification Of Functional Domain

FIGURE 1
Schematic representation of the workflow followed in this study.

Structural analysis

Physiochemical properties

The ExPASyProtparam tool was used to analyze the physiochemical properties such as molecular weight (Mwt), amino acid composition (AA), theoretical isoelectric point (pI), number of negative residues (–R), number of positive residues (+R), extinction coefficient (EC), half-life (h), instability index (II),aliphatic index (AI), and grand average of hydropathy (GRAVY) of the protein sequence (37).

Secondary structure prediction

The secondary structure of protein was predicted by using SOPMA and GOR IV. The self-optimized prediction method (SOPMA) describes the three states of the protein structure (helices, turns, and coils). SOPMA predicts 90% of secondary structural information of proteins and it works under the homologous method and predicts 69.5% of amino acids for three states of the secondary structure. SOPMA is mainly classified into four steps. Step one involves the retrieval of homologous protein from UniProt. In step two, alignments of sequence compose the set of homologous proteins. Step three executes the SOPMA method with each and every aligned sequence. In the final step, the conformational state yielding the highest score is attributed to the given amino acid with the averaged conformational score (14).

GOR IV (Garnier-Osguthorpe–Robson) is another method to predict the secondary structure. In version I, GOR has information from the hydrophobic triplet. Hydrophobic triplet information does not significantly improve the predictive power (15). The method GOR IV is formed on information theory; GOR has a mean accuracy of 64.4% for a three state prediction when compared to another version. Version IV is more accurate. The GOR IV method analyzes the secondary structure of the protein and correlates it with net values of each amino acid position and three states (helices, turns, coils) (16).

Tertiary structure prediction

The tertiary structure of vlhA genes was constructed using the homology modeling server RaptorX (http://raptorx.uchicago.edu/) and I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) (17). Raptor X distinguishes itself from other servers by the quality of the alignment between a target sequence and one or multiple distantly related template

TABLE 2 List of bioinformatics tools and servers employed in the structural and functional analyses of vlhA proteins.

S. no	Characterization/ analysis	Name of the server/tool	URL
1.	Phylogenetic analysis	Phylogeny.fr	http://www.phylogeny.fr/simple_phylogeny.cgi
2.	Physiochemical properties	ExPASy-Protparam tool	https://web.expasy.org/protparam/
3.	Secondary structure	SOPMA	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html
		GOR IV	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
4.	Tertiary structure	Raptor X	http://raptorx.uchicago.edu/
		I Tasser	https://zhanglab.dcmb.med.umich.edu/I-TASSER/
5.	Structure validation	PROCHECK	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html
		QMEAN	https://swissmodel.expasy.org/qmean/
6.	Sub cellular Localization	PSLPRED	http://crdd.osdd.net/raghava/pslpred/
		PSORTB	https://www.psort.org/psortb/
		CELLO2GO	http://cello.life.nctu.edu.tw/cello2go/
7.	Transmembrane Helix	SOSUI	https://harrier.nagahama-i-bio.ac.jp/sosui/mobile/
		HMMTOP	http://www.enzim.hu/hmmtop/
		TMHMM	http://www.cbs.dtu.dk/services/TMHMM/
8.	Signal peptide	Signal p	http://www.cbs.dtu.dk/services/SignalP/
		Target p	http://www.cbs.dtu.dk/services/TargetP/
9.	Functional Domain	CDD- BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
		HmmScan	https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan
		Pfam	http://pfam.xfam.org/
		SCANPROSITE	https://prosite.expasy.org/scanprosite/
		SMART	http://smart.embl-heidelberg.de/

proteins and by a novel nonlinear scoring function and a probabilistic-consistency algorithm. The predicted tertiary models can be used for binding site and epitope prediction; another application is found to be determining the binding topology of small ligand molecules to putative binding sites on the domain structure generated (54). The I-TASSER server employs *ab initio* modeling to predict 3D structures. The tertiary structures modeled by I-TASSER were subjected to refinement by the GalaxyRefine server (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE) (18). This server replaces amino acids with high-probability rotamers and applies molecular dynamic simulation for overall structural relaxation.

Structure validation

The refined structure was validated by PROCHECK (https://servicesn.mbi.ucla.edu/PROCHECK/), which analyzes the stereochemical quality of a protein structure by analyzing residue-by–Residue geometry and overall structure geometry (19).

QMEAN is used to analyze the quality of computationally predicted proteins. It is based on two distance-dependent interaction potentials of mean force, C-β atoms and is used to assess long–Range interactions (secondary structure dependent and torsion angle potential dependent). The QMEAN4 score is a linear combination of four statistical potential terms. It is trained to predict the IDDT (The Local Distance Difference Test) score in the range [0, 1]. To calculate the QMEAN Z-score, the normalized raw scores of a given model are compared with scores obtained for a representative set of high resolution X–Ray structures of similar size against the PDB reference set (20–22).

Functional analysis

Subcellular localization prediction

(A) PSLPRED

PSLpred is used for predicting the subcellular localization of prokaryotic proteins with an overall accuracy of 91.2%. It is a hybrid approach-based method. The prediction accuracies of 90.7, 86.8, 90.3, 95.2, and 90.6% were attained for cytoplasmic, extracellular, inner membrane, outer membrane, and periplasmic proteins, respectively (23).

(B) PSORTB

PSORTB is the most precise bacterial SCL (subcellular localization) prediction software that was introduced in 2005 and has been widely used. It provides quick and inexpensive means for gaining insight into the protein function, verifying experimental results, annotating newly sequenced bacterial genomes, detecting cell surface/drug targets, and identifying biomarkers for microbes. As a result, only ∼50% of proteins encoded in gram-negative bacterial genomes and ∼75% of proteins encoded in gram-positive bacterial genomes receive a prediction from PSORTb (24).

(C) CELLO2GO

CELLO2GO is a publicly available, web-based system for screening various properties of a targeted protein and its subcellular localization. It shows the exact location of the protein. CELLO2GO should be a useful tool for research involving complex subcellular systems because it combines CELLO and BLAST into one form (25).

Transmembrane helix prediction

(A) SOSUI

SOSUI is used for the discrimination of membrane proteins and soluble proteins and the prediction of the transmembrane helix, the accuracy of prediction was 99%, and the corresponding value for the transmembrane helix prediction was 97% (26).

(B) HMMTOP

A hidden Markov model with special architecture was developed to search transmembrane topology corresponding to the maximum likelihood among all the possible topologies of a given protein. The method is based on the hypothesis that the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins (27).

(C) TMHMM

TMHMM is a widely used bioinformatics tool, based on the hidden Markov model, which is used to predict transmembrane helices of integral membrane proteins. It is used to predict the number of transmembrane helices and discriminate between soluble and membrane proteins with a high degree of accuracy (28).

Signal peptide prediction

(A) Signal p

Signal p was the first publicly available method to predict signal peptide and its cleavage sites. It is based on deep neural network-based method combined with conditional random field classification and optimized transfer learning for improved signal peptide prediction. The input is given in FASTA format. The server predicts the presence of signal peptides, TAT signal peptides, and lipoprotein signal peptides from proteins present in Archaea, gram-positive bacteria, gram-negative bacteria, and eukaryotes (29).

(B) Target p

The target p server is used to predict the presence of signal peptides, and mitochondrial transit peptides and others were predicted using the FASTA sequence of the protein (30).

Identification of functional domain

The functional domain analysis was carried out using five publicly available tools (CDD-BLAST, HmmScan, Pfam, SCANPROSITE, and SMART). CDD-BLAST annotates the vlhA proteins by generating alignment models of the representative sequence fragment which were in agreement with domain boundaries as observed protein models in NCBI's Conserved Domain Database (31). HmmScan and SMART took a query sequence and searched it against the Pfam profile HMM library as a target database (32–34). Pfam was used to classify vlhA proteins functional families based on similarity (34). To predict the protein function, SCANPROSITE detects homologs and matches against signature from the PROSITE database (35).

Results

Phylogenetic analysis

Phylogenetic analysis was performed to examine the differences and relatedness among the vlhA proteins. A phylogenetic tree was constructed by using Phylogeny.fr. The computed data indicated that the expression of vlhA proteins during the course of infection varies greatly and vlhA from the five loci here clustered into different groups. The bootstrap values in the phylogenetic tree created for *M. gallisepticum* vlhA proteins showed that the proteins had less evolutionary similarity (Figure 2), and the divergence in sequence during evolution may have developed to evade host immune response and to adapt to each host. As a consequence, each protein has evolved due to strain during the course of infection, thus leading to antigenic variation (36).

Structural analysis

Physiochemical characterization

The ExPASy ProtParam was employed to analyze the protein primary structures and compute different parameters for their physiochemical properties. The number of amino acid residues in vlhA proteins varied from 77 to 795 amino acids. The

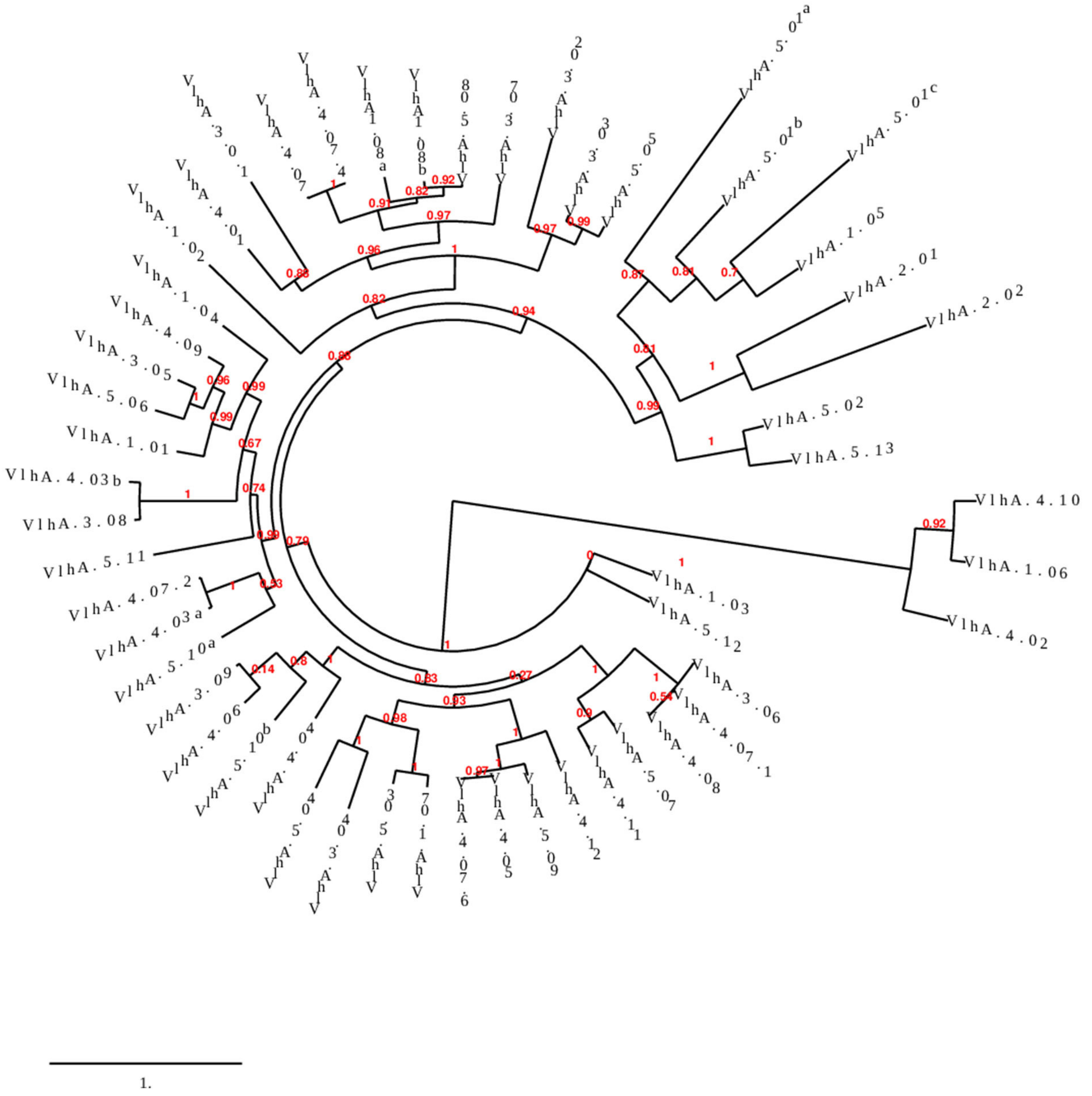


FIGURE 2
Phylogenetic tree showing the evolutionary relationship of different *M. gallisepticum* vlhA proteins. The numbers indicate bootstrap percentages and the scale indicates the divergence time.

composition of amino acid residues in each vlhA protein is presented in Figure 3. The molecular weight of these proteins varied from 8.12 to 85.3 kDa. The pI values of these proteins range from acidic pI 4.63 to alkaline pI 9.21. If the instability index (II) is above 40, the protein was considered to be unstable. As shown in Table 3, except a few vlhA (vlhA.1.08, vlhA.2.01, vlhA.5.01c, and vlhA.5.10b) proteins, other proteins were considerably stable. The aliphatic index (AI) of these vlhA proteins varied from 27.86 to 95.75. The high AI values indicated the thermal stability and hydrophobic nature of the proteins. When a protein was found to have a greater negative grand average of hydropathy (GRAVY) values, it indicated the hydrophilic nature of the protein and the possibility of better interactions between the protein and water (37). The complete physicochemical analysis of all the vlhA proteins is listed in Table 3.

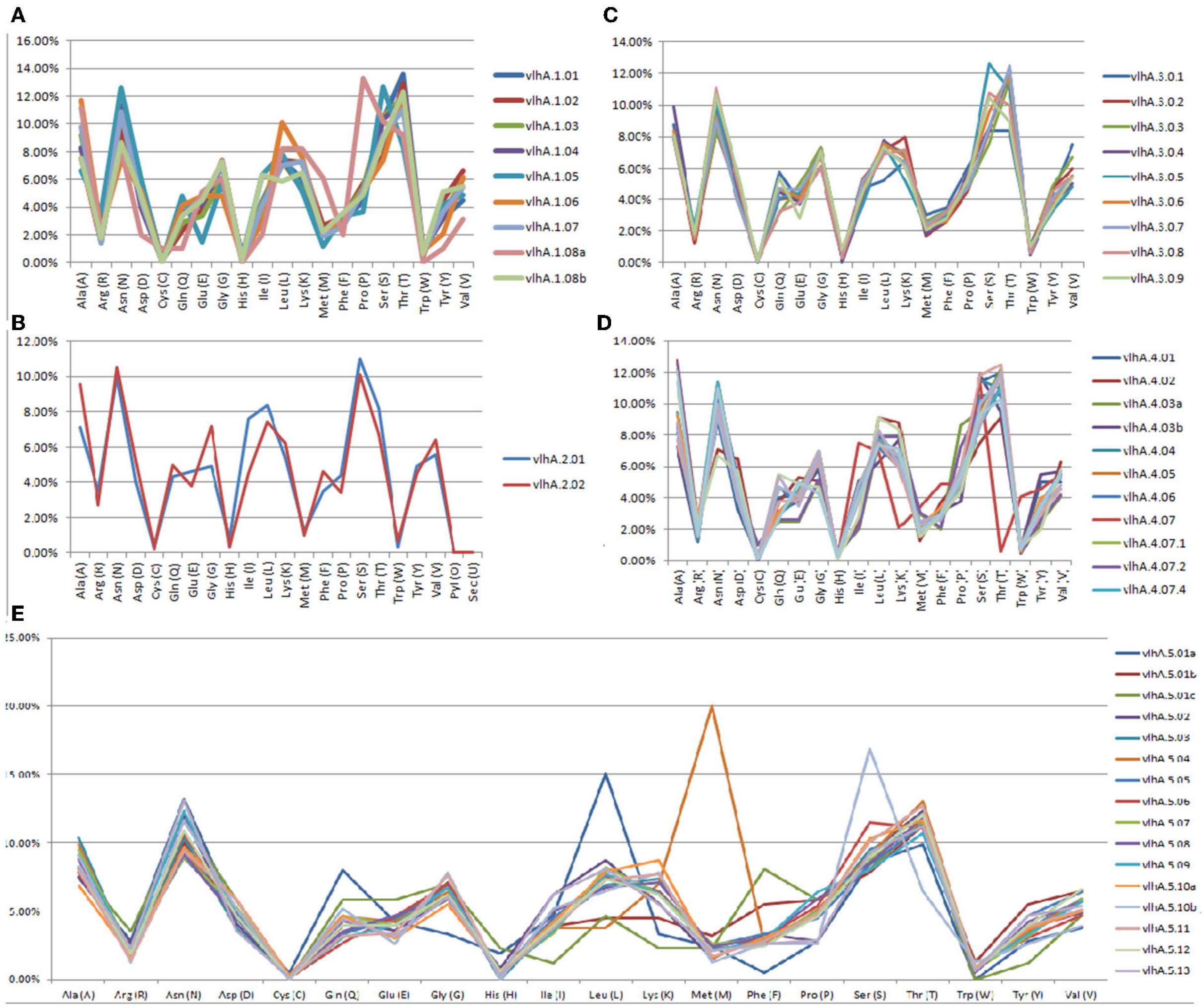


FIGURE 3
Graphical representation of amino acid composition of *M. gallisepticum* vlhA proteins. **(A)** vlhA 1 group, **(B)** vlhA 2 group, **(C)** vlhA 3 group, **(D)** vlhA 4 group, and **(E)** vlhA group 5.

Secondary structure prediction

The secondary structure of vlhA proteins was predicted using SOPMA and GOR IV servers that showed similar results where the percentage of random coils was higher when compared with alpha helices and extended turns (Supplementary Table 2). Previous studies reported that the presence of a higher percentage of random coil structures in bacterial proteins facilitated the dimerization and/or colocalization process and also act as adaptor proteins (38–43).

Three-dimensional structure modeling and validation

The tertiary models of vlhA proteins were constructed using the server called RaptorX and I Tasser. In tertiary models predicted by Raptor X, the number of amino acids was less compared to the input sequence, and thus the model predicted by I-TASSER was used for further analysis. The results from I-Tasser are consistent with the secondary structure prediction where these proteins were predicted to have a high percentage of random coil structures (Figure 4).

The PDBsum-PROCHECK program was used to validate the constructed three-dimensional models of these proteins. The Ramachandran Plot was used in the PROCHECK program to present the backbone conformation of proteins. The predicted models of vlhA proteins were analyzed and majority of the amino acid residues fall in the favored and allowed regions of the Ramachandran plot which indicates the good quality of the predicted models (Table 4).

QMEAN z-score was used to validate the good quality of these predicted tertiary models. This QMEAN software determined the closeness and similarity of the computationally

TABLE 3 Physiochemical properties like number of amino acids, molecular weight, isoelectric point, extinction coeffcient, half-life (h), instability index, aliphatic index, and GRAVY of *M. gallisepticum* vlhA proteins.

Protein name	Amino acid	Mol.wt	PI	Extinction coefficient	half Life (h)	instability Index	Aliphatic index	GRAVY
vlhA.1.01	686	74.02	6.23	71,280	30	26.86	67.46	−0.542
vlhA.1.02	666	71.65	5.3	61,200	30	25.89	68.83	−0.445
vlhA.1.03	682	72.83	5.54	63,260	30	28.14	71.85	−0.385
vlhA.1.04	697	74.83	6.81	65,780	30	32.57	68.75	−0.5
vlhA.1.05	730	79.70	9.08	73,690	30	36.44	36.44	−0.525
vlhA.1.06	754	80.92	6.36	62,340	30	27.55	80.44	−0.329
vlhA.1.07	728	77.55	5.49	67,730	30	30.03	69.08	−0.513
vlhA.1.08	98	10.21	9.25	1,490	30	46.91	59.9	−0.446
vlhA.1.08b	494	53.55	5.28	53,750	30	24.5	70.71	−0.414
vlhA.2.01	607	66.88	8.19	55,700	30	41.99	85.65	−0.354
vlhA.2.02	582	63.10	6.79	60,740	30	29.80	74.30	−0.430
vlhA.3.0.1	536	58.00	5.28	67,270	30	26.83	68.97	−0.442
vlhA.3.02	646	69.75	8.37	77,700	30	24.51	73.85	−0.439
vlhA.3.03	645	69.93	5.38	68,190	30	27.6	73.35	−0.389
vlhA.3.04	734	78.52	5.68	62,230	30	26.34	69.73	−0.515
vlhA.3.05	708	75.77	5.36	72,770	30	37.21	65.9	−0.531
vlhA.3.06	688	73.76	6.8	72,250	30	30.51	72.63	−0.427
vlhA.3.07	656	70.87	5.78	60,740	30	231.64	72.15	−0.426
vlhA.3.08	692	74.76	6	68,190	30	33.47	68.4	−0.537
vlhA.3.09	707	76.06	5.68	74,260	30	30.91	69.99	−0.55
vlhA.4.01	644	69.49	8.74	69,680	30	24.79	70.51	−0.415
vlhA.4.02	751	80.74	5.76	62,340	30	27.23	76.11	−0.438
vlhA.4.03a	197	20.71	9.06	13,075	30	24.19	61.57	−0.525
vlhA.4.03b	506	55.11	6.25	63,720	30	33.38	68.64	−0.523
vlhA.4.04	679	72.64	5.60	70,250	30	26.79	71.72	−0.465
vlhA.4.05	673	72.27	6.01	67,270	30	25.63	71.66	−0.466
vlhA.4.06	698	74.98	5.56	74,260	30	30.59	66.85	−0.545
vlhA.4.07	667	71.81	8.72	62,230	30	30.34	67.80	−0.501
vlhA.4.07.1	684	73.2	7.56	70,250	30	30.66	72.35	−0.424
vlhA.4.07.2	191	20.1	9.21	13,075	30	24.54	63.51	−0.493
vlhA.4.07.4	673	72.3	6.32	68,760	30	25.37	71.66	−0.463
vlhA.4.07.6	667	71.7	8.30	62,230	30	29.85	67.80	−0.496
vlhA.4.08	688	73.6	7.56	70,250	30	30.54	71.93	−0.428
vlhA.4.09	710	76	6.88	69,790	30	31.84	65.17	−0.514
vlhA.4.10	795	85.3	7.52	62,340	30	30.29	74.34	−0.479
vlhA4.11	690	74	6.40	61,770	30	28.52	65.82	−0.544
vlhA.4.12	701	75.1	5.28	63,260	30	27.86	27.86	−0.446
vlhA.5.01a	212	23.32	5.10	8,940	30	38.68	95.75	−0.456
vlhA.5.01b	309	33.93	4.80	47,330	30	31.19	59.35	−0.431
vlhA.5.01c	86	9.38	4.63	1,490	30	41.87	44.30	−0.779
vlhA.5.02	610	66.45	8.51	56,270	30	30.65	79.98	−0.407
vlhA.5.03	728	77.47	8.78	67,730	30	28.91	72.15	−0.449
vlhA.5.04	740	78.90	5.17	66,240	30	35.27	69.27	−0.467
vlhA.5.05	644	69.83	5.73	66,700	30	26.68	73.93	−0.392

(Continued)

TABLE 3 Continued

Protein name	Amino acid	Mol.wt	PI	Extinction coefficient	half Life (h)	instability Index	Aliphatic index	GRAVY
vlhA.5.06	703	75.28	5.75	65,780	30	28.52	65.82	−0.544
vlhA.5.07	681	73.27	5.55	60,280	30	30.09	71.82	−0.444
vlhA.5.08	661	71.41	6.32	58,220	30	29.53	70.88	−0.460
vlhA.5.09	701	75.19	6.42	67,270	30	24.61	69.83	−0.531
vlhA.5.10a	642	70.06	9.04	68,885	30	26.23	68.69	−0.619
vlhA.5.10b	77	8.12	8.03	8,480	30	51.99	65.97	−0.619
vlhA.5.11	711	75.88	6.87	69,220	30	20.48	66.03	−0.55
vlhA.5.12	678	73.12	5.81	67,730	30	25.23	71.80	−0.483
vlhA.5.13	616	66.94	8.89	65,210	30	29.03	79.53	−0.394

predicted model with the existing PDB reference set. The normalized QMEAN score is provided in Table 4.

Functional analysis

Localization of vlhA proteins

In this study, 3 different servers (CELLO2GO, PSORTB, and PSLPRED) were used to predict the cellular location of vlhA proteins. As provided in Table 4, the vlhA proteins were predicted to be extracellular proteins which help in the host interactions and immune evasion. The results were similar for all the three servers. TMMHMM, HMMTOP, and SOSUI servers were used to predict the presence of transmembrane helices in these proteins. Except vlhA-—1.08b, 2.01, 2.02, 3.01, 3.02, 3.08, 4.01, 4.03b, 5.01a, 5.01b, 5.01c, 5.02, 5.08, 5.10b, and 5.13, other proteins were predicted to have transmembrane helices (Table 4). The prediction results are consistent among the servers. Based on the prediction using SignalP and TargetP servers, several vlhA proteins having lower values indicated the absence of signal peptides in them. In contrast, the vlhA proteins with higher values indicated the presence of signal peptides in their sequence (Table 4).

Identification of the functional domain

There are a large number of proteins that have no assigned function. For those proteins, the annotation generally depends on the sequence homology techniques (21). Functional domains were identified using CDD- BLAST, HmmScan, Pfam, SCANPROSITE, and SMART publicly available tools. After screening the vlhA proteins in the above mentioned servers, all the proteins were grouped under the mycoplasma hemagglutinin family by all the servers. Based on the similarity of the sequences of these proteins with mycoplasma hemagglutinin, these proteins were predicted to play a role in the hemagglutination process. The mycoplasma hemagglutinin family consists of several hemagglutinin sequences from mycoplasma species. The major plasma membrane proteins, vlhAs, of *M. gallisepticum* are cell adhesions or hemagglutinin molecules. The hemagglutination process of mycoplasma plays a crucial role in host immune evasion; the exact mechanism through which the hemagglutination mediated immune evasion occurs is yet to be explored (44, 45).

Discussion

Variable lipoprotein hemagglutinin A gene encodes immunodominant proteins that are believed to be responsible for *M. gallisepticum*'s host cell interaction, pathogenesis, and immune evasion; however, their exact mechanism is unknown (46). The sound knowledge about the mechanism of immune evasion by this protein family will be valuable in the development of drugs and vaccines against *M. gallisepticum* infection in chickens. Protein structure and function identification is an essential step for understanding its cellular and molecular processes. *In silico* homology modeling studies provide an opportunity to establish a route for the structural modeling and analysis of vlhA proteins. With rapid advances in bioinformatics and computational biology, the prediction and validation of the structure and function of proteins have become easily accessible. The importance of functional analysis of proteins includes deeper knowledge in molecular mechanisms of disease progression, exploration of effective prophylactic targets, relationship, and interaction with other proteins in the same microorganism.

This study has analyzed the vlhA proteins from *M. gallisepticum* strain R low for its structural and functional characteristics. The amino acid sequences of vlhA proteins were retrieved in FASTA format from the UniProt database and used for further structural and functional analyses. The physiochemical characteristics such as amino acid composition, isoelectric point (pI), number of negative and positive residues,

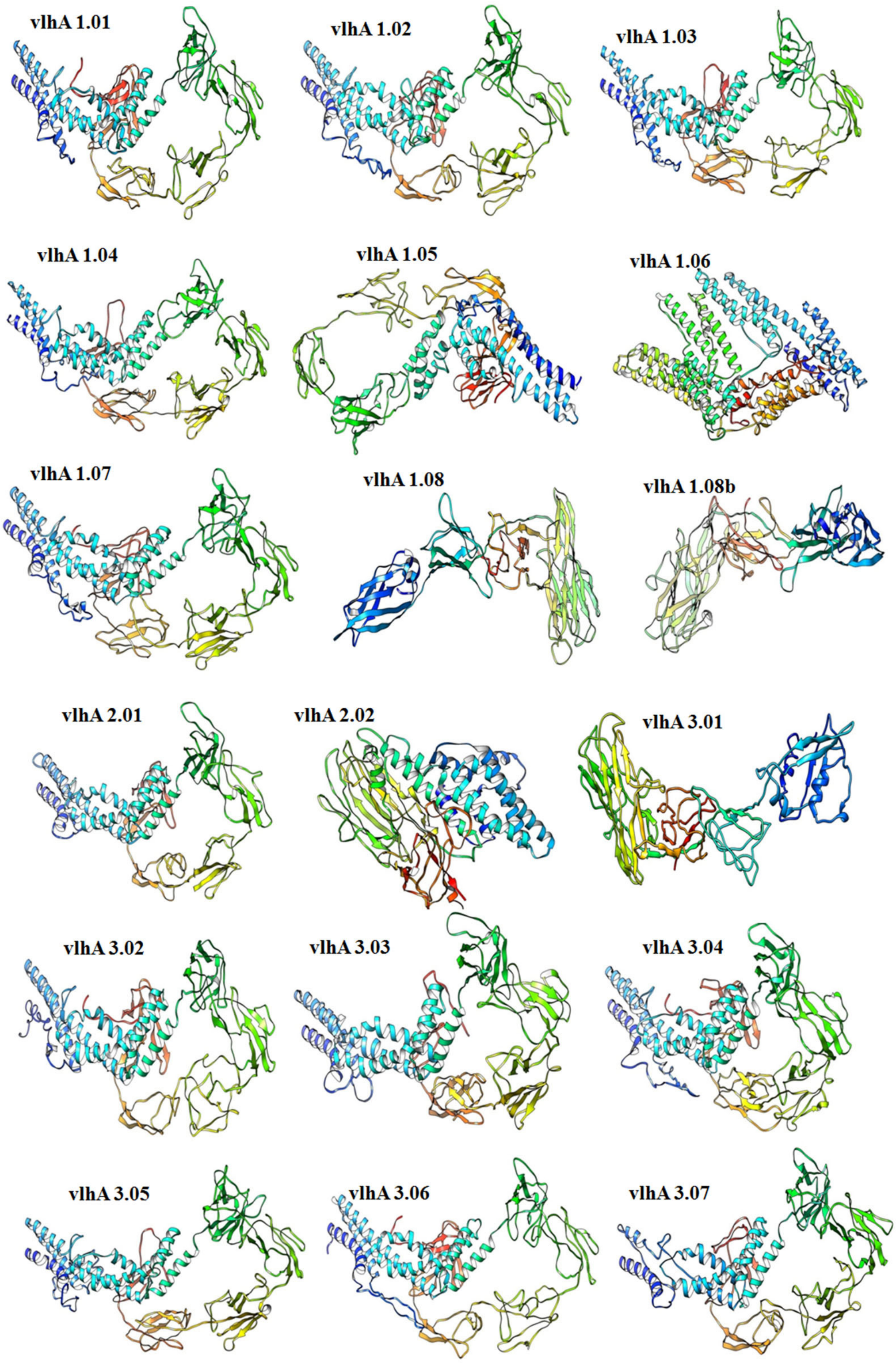


FIGURE 4
Continued.

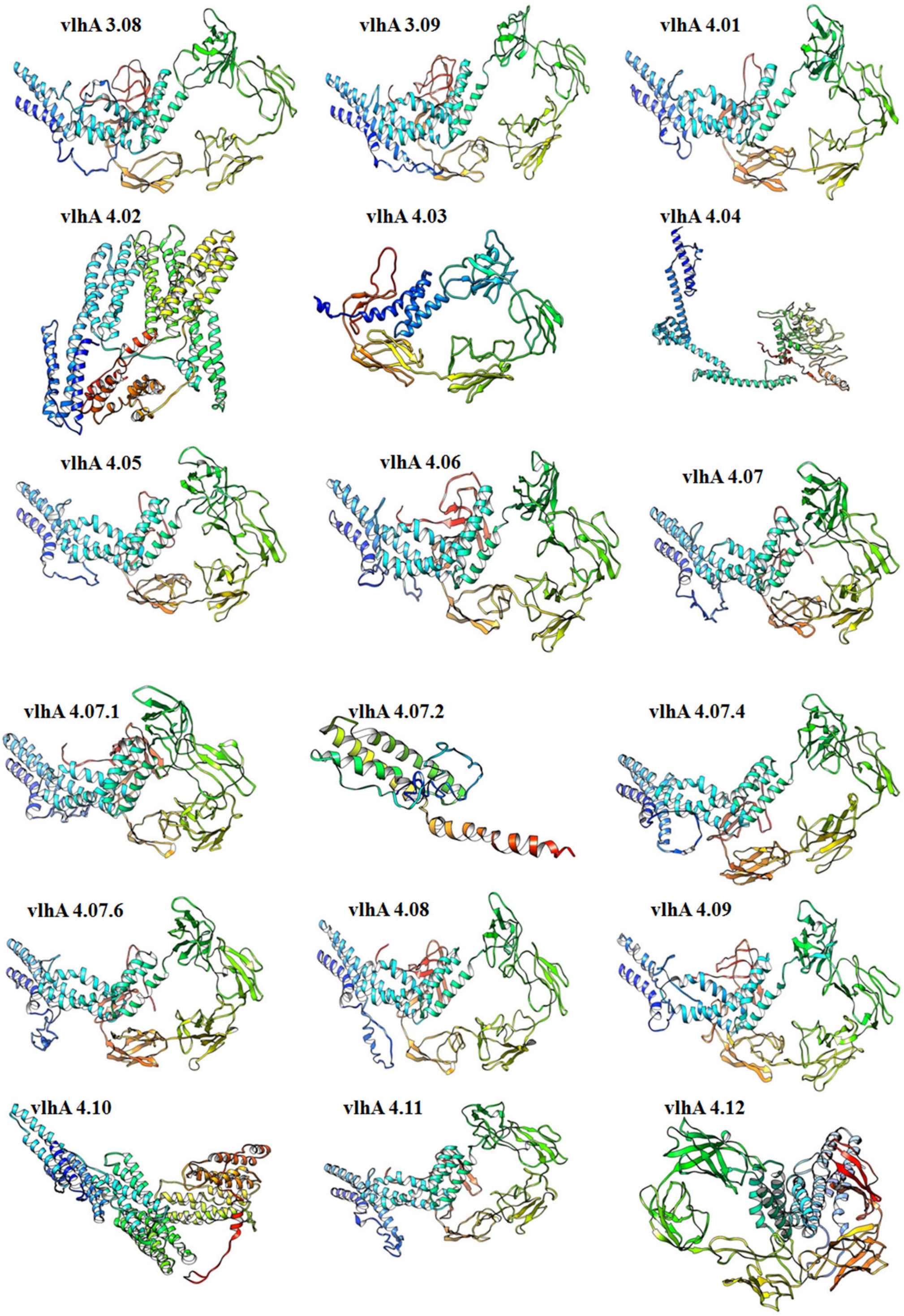


FIGURE 4
Continued.

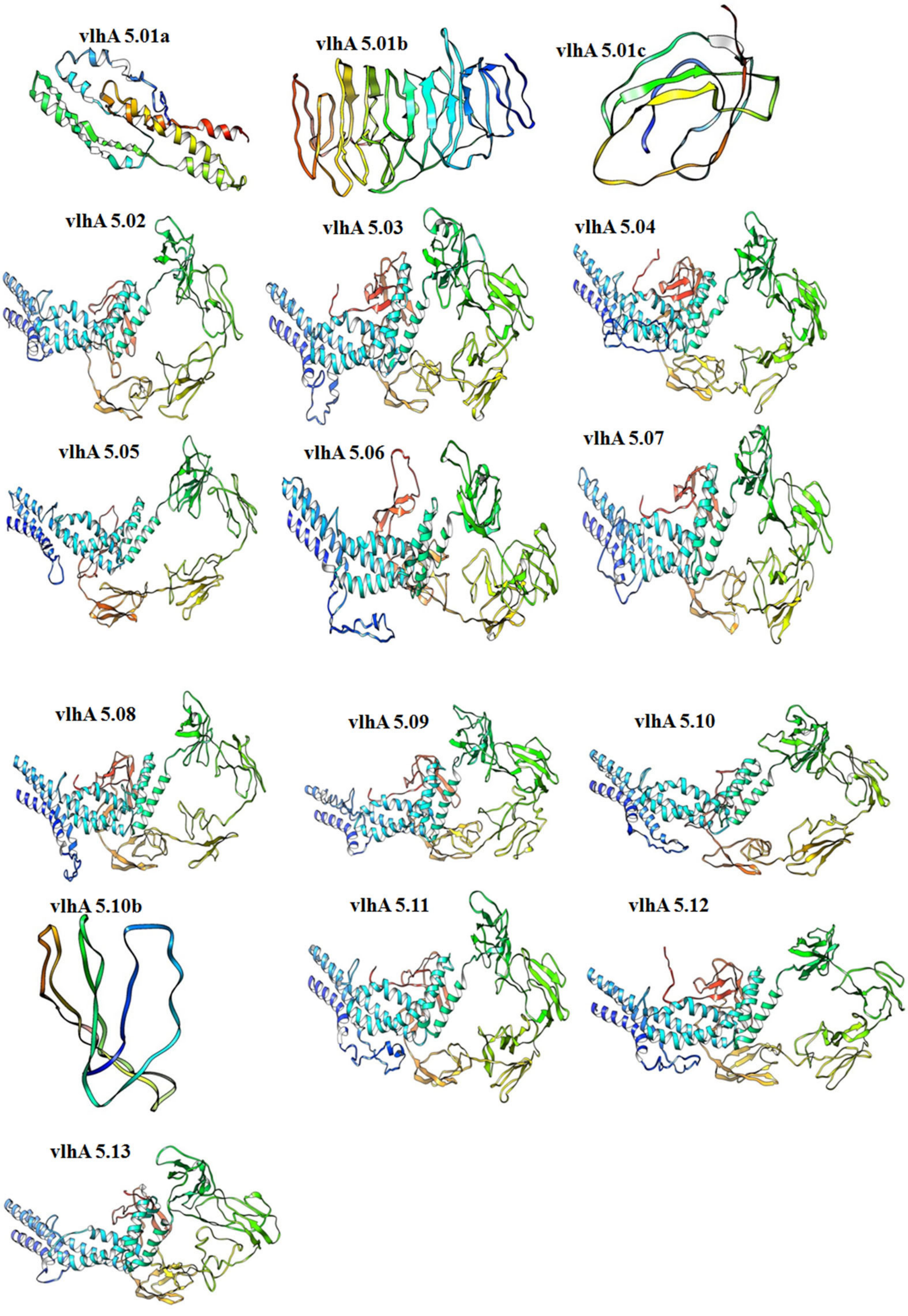


FIGURE 4
Three-dimensional *ab initio* models of vlhA proteins. Visualizations of model structures were performed by UCSF Chimera.

extinction coefficient, half-life, instability index (II), aliphatic index (AI), and grand average of hydropathy (GRAVY) of these proteins were predicted. According to the results obtained, a higher number of amino acids such as threonine, asparagine, serine, and alanine were observed whereas the amino acids such as cysteine, histidine, and tryptophan were low in amount.

Cysteines are important for the formation of disulfide bonds in the protein structure which cannot be easily substituted or replaced and often acts together with histidines which are commonly present in the active or binding sites of the proteins (38). These vlhA proteins have the average molecular weight of 59.28 kDa, and are hydrophilic in nature and stable.

TABLE 4 Tertiary structural validation- Qmean Score, Ramachandran plot most favored region and functional analysis-Subcellular Localization, Transmembrane helix, Signal peptide of vlhA proteins.

S.No	Protein name	Qmean score	Ramachandran plot most favored region	Subcellular localization	Transmenbrane helix	Signal peptide
1	vlhA.1.01	−10.78	68.1%	Extracellular	2(44–61)(106–123)	Yes
2	vlhA.1.02	−10.06	70.2%	Extracellular	2(42–59)(104–121)	Yes
3	vlhA.1.03	−9.32	69.3%	Extracellular	2(42–59)(104–121)	Yes
4	vlhA.1.04	−10.62	69.6%	Extracellular	2(42–59)(104–121)	Yes
5	vlhA.1.05	−11.63	66.3%	Extracellular	0	Yes
6	vlhA.1.06	−7.45	82.0%	Periplasmic	2(42–59)(104–121)	Yes
7	vlhA.1.07	−10.65	68.4%	Extracellular	2(42–59)(104–121)	No
8	vlhA.1.08	−12.84	65.5%	Periplasmic	2(66–83)(126–146)	Yes
9	vlhA.1.08b	−12.84	65.5%	Extracellular	0	Yes
10	vlhA.2.01	−10.16	66.4%	Extracellular	0	Yes
11	vlhA.2.02	−9.48	69.0%	Extracellular	2(42–59)(104–121)	No
12	vlhA.3.0.1	−14.31	52.8%	Periplasmic	0	Yes
13	vlhA.3.02	−9.78	67.1%	Extracellular	0	Yes
14	vlhA.3.03	−10.31	68.0%	Extracellular	2(46–63)(108–125)	No
15	vlhA.3.04	−10.86	66.4%	Extracellular	2(46–63)(108–125)	Yes
16	vlhA.3.05	−10.04	69.3%	Extracellular	2(46–63)(108–125)	Yes
17	vlhA.3.06	−11.47	65.8%	Extracellular	1(9–26)	Yes
18	vlhA.3.07	−11.57	62.5%	Extracellular	1(9–26)	Yes
19	vlhA.3.08	−10.78	66.5%	Extracellular	1(9–26)	Yes
20	vlhA.3.09	−9.09	66. 2%	Extracellular	1(9–26)	Yes
21	vlhA.4.01	−9.02	69.1%	Extracellular	1(9–26)	No
22	vlhA.4.02	−7.77	81.5%	Periplasmic	1(9–26)	Yes
23	vlhA.4.03a	−10.23	66.1%	Outermenbrane	1(9–26)	Yes
24	vlhA.4.03b	−12.26	66.3%	Extracellular	0	Yes
25	vlhA.4.04	−12.23	60.7%	Extracellular	2(44–61)(106–123)	Yes
26	vlhA.4.05	−10.63	66.3%	Extracellular	2(44–61)(106–123)	Yes
27	vlhA.4.06	−10.28	67.5%	Extracellular	2(44–61)(106–123)	Yes
28	vlhA.4.07	−10.81	66.7%	Extracellular	2(44–61)(106–123)	Yes
29	vlhA.4.07.1	−11.19	67.6%	Extracellular	2(46–63) (108–125)	Yes
30	vlhA.4.07.2	−11.85	60.6%	Extracellular	2(66–83) (129–146)	Yes
31	vlhA.4.07.4	−9.15	68. 2%	Extracellular	2(46–63) (108–125)	Yes
32	vlhA.4.07.6	−10.71	67.4%	Extracellular	2(46–63) (108–125)	Yes
33	vlhA.4.08	−10.65	65.5%	Extracellular	3(10–27) (44–61) (106–123)	Yes
34	vlhA.4.09	−11.45	66.5%	Outermenbrane	2(44–61) (106–123)	Yes
35	vlhA.4.10	−7.86	81.4%	Periplasmic	2(44–61) (106–123)	Yes
36	vlhA4.11	−10.76	65.6%	Extracellular	2(44–61) (106–123)	Yes
37	vlhA.4.12	−10.43	67.9%	Outermenbrane	2(44–61) (106–123)	Yes
38	vlhA.5.01a	−12.07	56.3%	Extracellular	0	Yes
39	vlhA.5.01b	−13.51	40.4%	Extracellular	0	Yes
40	vlhA.5.01c	−8.69	38.4%	Extracellular	0	Yes

(Continued)

TABLE 4 Continued

S.No	Protein name	Qmean score	Ramachandran plot most favored region	Subcellular localization	Transmenbrane helix	Signal peptide
41	vlhA.5.02	−10.15	67.0%	Extracellular	0	Yes
42	vlhA.5.03	−11.14	69.7%	Extracellular	2(44–61)	Yes
43	vlhA.5.04	−10.63	70.0%	Extracellular	2(44–61)(106–123)	Yes
44	vlhA.5.05	−9.17	69.0%	Extracellular	2(44–61)(106–123)	Yes
45	vlhA.5.06	−9.65	67.1%	Extracellular	1 (19–38)	Yes
46	vlhA.5.07	−10.58	67.1%	Extracellular	2(44–61) (106–123)	Yes
47	vlhA.5.08	−12.05	66.6%	Extracellular	2(44–61) (106–123)	Yes
48	vlhA.5.09	−11.32	66.4%	Extracellular	2(44–61) (106–123)	Yes
49	vlhA.5.10a	−9.33	70.1%	Extracellular	2(64–81)(127–144)	Yes
50	vlhA.5.10b	−11.02	25.4%	Outermenbrane	0	Yes
51	vlhA.5.11	−11.26	67.6%	Extracellular	2(44–61)(106–123)	Yes
52	vlhA.5.12	−11.73	71.3%	Extracellular	2(44–61)(106–123)	Yes
53	vlhA.5.13	−10.82	66.4%	Extracellular	0	Yes

The secondary structure of these proteins contains a higher percentage of random coils which are believed to facilitate in the dimerization and/or colocalization process and may also act as adaptor proteins (39–43, 53). The tertiary structures of vlhA proteins were predicted and validated for the good quality of the computationally predicted protein structure. These proteins have been predicted to be stable with the higher percentage of amino acids present in the most favored regions (>80%). The obtained QMEAN score indicated the good quality of these proteins with higher QMEAN values (20). As for the functional prediction of vlhA proteins, all of these proteins were predicted to be extracellular which may subsequently help in the immune evasion of the *M. gallisepticum* from the host immune system. The identification of the functional domain was performed by the sequence homology techniques. The result obtained showed that the domains of these proteins were similar to the mycoplasma hemagglutinin family as they consist of hemagglutinin sequences from the mycoplasma family and predicted to be involved in the hemagglutination process. It has been reported that the genetic determinants that code for the hemagglutinins are organized into a large family of genes and that only one of these genes is predominately expressed during the course of infection at a given time (44, 47–49). Antigenic variation or phenotypic switching occurs due to high frequency genetic mutations. Due to the lack of a rigid cell wall, the lipoproteins in the mycoplasma cell membrane function as the major elements that come into contact with the host environment (45, 46, 50). These proteins undergo antigenic variation through on/off switching, domain shuffling, and size variation to modify the antigenic components on their cell surface to produce heterotypes that allow mycoplasma to evade recognition and clearance by host immune cells that largely eliminate homo-types. Numerous human and animal mycoplasma species have the ability to go through antigenic variation so that these bacteria can evade recognition by the host humoral immune system (51, 52). In *M. gallisepticum,* the hemagglutination process may play a role in triggering the antigenic variation cascade leading to immune evasion. Since the exact function and machinery of these vlhA proteins are not determined at present, the *in silico* structural and functional prediction of these proteins may help in the determination of its cellular and molecular processes. To the best of our knowledge, this is the first study to explore the structural and functional properties of vlhA proteins. These findings may aid in understanding the mechanism of immune evasion by vlhA proteins.

Conclusion

Identifying the molecular processes by which the vlhA protein evades the host immune response is critical in understanding the pathogenicity of *M. gallisepticum* and will aid in the development of efficient infection control measures. *In silico* homology modeling studies allow researchers to build a pipeline for structural modeling and functional analysis of any protein as part of discovering the molecular mechanism of the protein's function and therapeutic targets. The physicochemical features of selected vlhA that are important for immune evasion were given in this work. The study also included secondary structure and tertiary model characteristics for the vlhA proteins. Furthermore, the functional analysis revealed that the vlhA proteins are clustered under the mycoplasma hemagglutinin family. For functional analysis of vlhA proteins, multiple servers like CDD- BLAST, HmmScan, Pfam, SCANPROSITE, and SMART were used and all the servers grouped the vlhA proteins under the mycoplasma hemagglutinin family; the results obtained were consistent, thus validating the uniqueness of our findings. The significance of this study is the analysis and exploration of unknown structural and functional characteristics of vlhA proteins through the application of latest bioinformatics software like Protparam, I Tasser, PSORTB, TMMHMM, SignalP, and Pfam, thus bridging the gap in knowledge in the role of vlhA proteins in *M. gallisepticum* pathogenesis. This research will serve as a foundation for future experimental studies aimed at clarifying the functional molecular mechanism of immune response.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

SM and MH designed and performed the experimental studies. SM carried out the *in silico* experiments. The manuscript was written by SM and MH. Both authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.943831/full#supplementary-material

References

1. Ley DH. *Mycoplasma gallisepticum* infection. In:. Saif YM, Fadly AM, Glisson JR, McDougald LR, Nolan LK, Swayne DE, eds *Diseases of poultry, 12 ed.* Ames, IA: Blackwell Publishing Professional (2008). 807–45.

2. Winner F, Rosengarten R, Citti C. In vitro cell invasion of *Mycoplasma gallisepticum. Infect Immun.* (2000) 68:4238–44. doi: 10.1128/IAI.68.7.4238-4244.2000

3. Tulman ER, Liao X, Szczepanek SM, Ley DH, Kutish GF, Geary SJ, et al. Extensive variation in surface lipoprotein gene content and genomic changes associated with virulence during evolution of a novel North American house finch epizootic strain of *Mycoplasma gallisepticum*. *Microbiol.* (2012) 158:2073–88. doi: 10.1099/mic.0.058560-0

4. Pflaum K, Tulman ER, Beaudet J, Liao X, Geary SJ. Global changes in *Mycoplasma gallisepticum* phase-variable lipoprotein gene vlha expression during *in vivo* infection of the natural chicken host. *Infect Immun.* (2015) 84:351–5. doi: 10.1128/IAI.01092-15

5. Noormohammadi AH. Role of phenotypic diversity in pathogenesis of avian mycoplasmosis. *Avian Pathol.* (2007) 36:439–44. doi: 10.1080/03079450701687078

6. Glew MD, Browning GF, Markham PF, Walker ID. pMGA phenotypic variation in *Mycoplasma gallisepticum* occurs *in vivo* and is mediated by trinucleotide repeat length variation. *Infect Immun.* (2000) 68:6027–33. doi: 10.1128/IAI.68.10.6027-6033.2000

7. Chopra-Dewasthaly R, Spergser J, Zimmermann M, Citti C, Jechlinger W, Rosengarten R, et al. Vpma phase variation is important for survival and persistence of Mycoplasma agalactiae in the immunocompetent host. *PLoS Pathog.* (2017) 13:e1006656. doi: 10.1371/journal.ppat.1006656

8. Czurda S, Hegde SM, Rosengarten R, Chopra-Dewasthaly R. Xer1-independent mechanisms of Vpma phase variation in Mycoplasma agalactiae are triggered by Vpma-specific antibodies. *Int J Med Microbiol.* (2017) 307:443–51. doi: 10.1016/j.ijmm.2017.10.005

9. Ma L, Jensen JS, Mancuso M, Myers L, Martin DH. Kinetics of Genetic Variation of the *Mycoplasma genitalium* MG192 gene in experimentally infected chimpanzees. *Infect Immun.* (2015) 84:747–53. doi: 10.1128/IAI.01162-15

10. Citti C, Nouvel LX, Baranowski E. Phase and antigenic variation in mycoplasmas. *Future Microbiol.* (2010) 5:1073–85. doi: 10.2217/fmb.10.71

11. Hu F, Zhao C, Bi D, Tian W, Chen J, Sun J, et al. *Mycoplasma gallisepticum* (HS strain) surface lipoprotein pMGA interacts with host apolipoprotein A-I during infection in chicken. *Appl Microbial Biotechnol.* (2016) 100:1343–54. doi: 10.1007/s00253-015-7117-9

12. Vogl G, Plaickner A, Szathmary S, Stipkovits L, Rosengarten R, Szostak MP, et al. *Mycoplasma gallisepticum* invades chicken erythrocytes during infection. *Infect Immun.* (2008) 76:71–7. doi: 10.1128/IAI.00871-07

13. Dereeper A, GuignonV, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* (2008) 36:W465–9. doi: 10.1093/nar/gkn180

14. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci.* (1995) 11:681–4. doi: 10.1093/bioinformatics/11.6.681

15. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol.* (1978) 120:97–120. doi: 10.1016/0022-2836(78)90297-8

16. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* (1996) 266:540–53. doi: 10.1016/S0076-6879(96)66034-0

17. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y, et al. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* (2015) 12:7–8. doi: 10.1038/nmeth.3213

18. Heo L, Park H, Seok C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* (2013) 41:W384–8. doi: 10.1093/nar/gkt458

19. Laskowski RA, MacArthur MW, Moss DS, Thornton J. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* (1993) 26:283–91. doi: 10.1107/S0021889892009944

20. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinform.* (2011) 27:343–50. doi: 10.1093/bioinformatics/btq662

21. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinform.* (2013) Chapter 3:Unit3.1. doi: 10.1002/0471250953.bi0301s42

22. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinform.* (2013) 29:2722–8. doi: 10.1093/bioinformatics/btt473

23. Bhasin M, Garg A, Raghava GP. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinform.* (2005) 21:2522–4. doi: 10.1093/bioinformatics/bti309

24. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinform.* (2010) 26:1608–1615. doi: 10.1093/bioinformatics/btq249

25. Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK, et al. CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PLoS ONE.* (2014) 9:e99368. doi: 10.1371/journal.pone.0099368

26. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinform.* (1998) 14:378–9. doi: 10.1093/bioinformatics/14.4.378

27. Tusnády GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol.* (1998) 283:489–506. doi: 10.1006/jmbi.1998.2107

28. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* (2001) 305:567–80. doi: 10.1006/jmbi.2000.4315

29. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* (2019) 37:420–23. doi: 10.1038/s41587-019-0036-z

30. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance.* (2019) 2:e201900429. doi: 10.26508/lsa.201900429

31. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* (2015) 43:D222–6. doi: 10.1093/nar/gku1221

32. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* (2011) 39:W29–37. doi: 10.1093/nar/gkr367

33. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* (2000) 28:231–4. doi: 10.1093/nar/28.1.231

34. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res.* (2014) 42:D222–30. doi: 10.1093/nar/gkt1223

35. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* (2006) 34:W362–5. doi: 10.1093/nar/gkl124

36. Bencina D, Drobnic-Valic M, Horvat S, Narat M, Kleven SH, Dovc P, et al. Molecular basis of the length variation in the N-terminal part of Mycoplasma synoviae hemagglutinin. *FEMS Microbiol Lett.* (2001) 203:115–23. doi: 10.1111/j.1574-6968.2001.tb10829.x

37. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A, et al. Protein identification and analysis tools on the ExPASy server. *The Proteomics Protocols Handbook.* Berlin/Heidelberg, Germany: Springer. (2005). p. 571–607.

38. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. et al. *Molecular Biology of the Cell. 4th edition.* New York: Garland Science (2002). Available from: https://www.ncbi.nlm.nih.gov/books/NBK26830/

39. Alaidarous M, Ve T, Casey LW, Valkov E, Ericsson DJ, Ullah MO, et al. Mechanism of bacterial interference with TLR4 signaling by Brucella Toll/interleukin-1 receptor domain-containing protein TcpB. *J Biol Chem.* (2014) 289:654–68. doi: 10.1074/jbc.M113.523274

40. Xiong D, Song L, Geng S, Jiao Y, Zhou X, Song H, et al. Salmonella coiled-coil- and TIR-containing TcpS evades the innate immune system and subdues inflammation. *Cell Rep.* (2019) 28, 804–818.e7. doi: 10.1016/j.celrep.2019.06.048

41. Radhakrishnan GK, Yu Q, Harms JS, Splitter GA. Brucella TIR domain-containing protein mimics properties of the toll-like receptor adaptor protein TIRAP. *J Biol Chem.* (2009) 284:9892–8. doi: 10.1074/jbc.M805458200

42. Rana RR, Zhang M, Spear AM, Atkins HS, Byrne B. Bacterial TIR-containing proteins and host innate immune system evasion. *Med Microbiol Immunol.* (2013) 202:1–10. doi: 10.1007/s00430-012-0253-2

43. Ve T, Gay NJ, Mansell A, Kobe B, Kellie S. Adaptors in toll-like receptor signaling and their potential as therapeutic targets. *Curr Drug Targets.* (2012) 13:1360–74. doi: 10.2174/138945012803530260

44. Rosengarten R, Citti C, Glew M, Lischewski A, Droesse M, Much P, et al. Host-pathogen interactions in mycoplasma pathogenesis: virulence and survival strategies of minimalist prokaryotes. *Int J Med Microbiol.* (2000) 290:15–25. doi: 10.1016/S1438-4221(00)80099-5

45. Christodoulides A, Gupta N, Yacoubian V, Maithel N, Parker J, Kelesidis T, et al. The role of lipoproteins in mycoplasma-mediated immunomodulation. *Front Microbiol.* (2018) 9:1682. doi: 10.3389/fmicb.2018.01682

46. Pflaum K, Tulman ER, Beaudet J, Canter J, Geary SJ. Variable Lipoprotein Hemagglutinin A Gene (vlhA) expression in variant *Mycoplasma gallisepticum* strains *in vivo*. *Infect Immun.* (2018) 86:e00524–18. doi: 10.1128/IAI.00524-18

47. Liu L, Payne DM, van Santen VL, Dybvig K, Panangala VS. A protein (M9) associated with monoclonal antibody-mediated agglutination of *Mycoplasma gallisepticum* is a member of the pMGA family. *Infect Immun.* (1998) 66:5570–5. doi: 10.1128/IAI.66.11.5570-5575.1998

48. Markham PF, Glew MD, Sykes JE, Bowden TR, Pollocks TD, Browning GF, et al. The organisation of the multigene family which encodes the major cell surface protein, pMGA, of *Mycoplasma gallisepticum*. *FEBS Lett.* (1994) 352:347–52. doi: 10.1016/0014-5793(94)00991-0

49. Noormohammadi AH, Markham PF, Kanci A, Whithear KG, Browning GF. A novel mechanism for control of antigenic variation in the haemagglutinin gene family of mycoplasma synoviae. *Mol Microbial.* (2000) 35:911–23. doi: 10.1046/j.1365-2958.2000.01766.x

50. Barbosa MS, Spergser J, Marques LM, Timenetsky J, Rosengarten R, Chopra-Dewasthaly R, et al. Predominant single stable VpmaV expression in strain GM139 and major differences with *Mycoplasma agalactiae* type strain PG2. *Animals.* (2022) 12:265. doi: 10.3390/ani12030265

51. Pflaum K, Tulman ER, Canter J, Dhondt KV, Reinoso-Perez MT, Dhondt AA, et al. The influence of host tissue on *M. gallisepticum* vlhA gene expression. *Vet Microbiol.* (2020) 251:108891. doi: 10.1016/j.vetmic.2020.108891

52. Galyamina MA, Zubov AI, Ladygina VG, Li AV, Matyushkina DS, Pobeguts OV, et al. Comparative proteomic analysis of the *Mycoplasma gallisepticum* nucleoid fraction before and after infection. *Bull Exp Biol Med.* (2022) 172:336–40. doi: 10.1007/s10517-022-05388-4

53. Alaidarous M. In silico structural homology modeling and characterization of multiple N-terminal domains of selected bacterial Tcps. *PeerJ.* (2020) 8.e10143. doi: 10.7717/peerj.10143

54. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. *Nat Protocol.* (2012) 7:1511–22. doi: 10.1038/nprot.2012.085

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fvets.2022.1010039

OPEN ACCESS

EDITED BY
Jun-Mo Kim,
Chung-Ang University, South Korea

REVIEWED BY
Nouman Amjad,
Islamia University of Bahawalpur, Pakistan
Tansol Park,
Chung-Ang University, South Korea

*CORRESPONDENCE
Matthew A. Scott
mascott@cvm.tamu.edu

[†]These authors share senior authorship

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

RECEIVED 02 August 2022
ACCEPTED 25 August 2022
PUBLISHED 26 September 2022

CITATION
Scott MA, Woolums AR, Karisch BB, Harvey KM and Capik SF (2022) Impact of preweaning vaccination on host gene expression and antibody titers in healthy beef calves. *Front. Vet. Sci.* 9:1010039. doi: 10.3389/fvets.2022.1010039

Impact of preweaning vaccination on host gene expression and antibody titers in healthy beef calves

Matthew A. Scott[1*†], Amelia R. Woolums[2], Brandi B. Karisch[3], Kelsey M. Harvey[4] and Sarah F. Capik[5,6†]

[1]Veterinary Education, Research, and Outreach Center, Texas A&M University and West Texas A&M University, Canyon, TX, United States, [2]Department of Pathobiology and Population Medicine, Mississippi State University, Mississippi State, MS, United States, [3]Department of Animal and Dairy Sciences, Mississippi State University, Mississippi State, MS, United States, [4]Prairie Research Unit, Mississippi State University, Prairie, MS, United States, [5]Texas A&M AgriLife Research, Texas A&M University System, Amarillo, TX, United States, [6]Department of Veterinary Pathobiology, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, United States

The impact of preweaning vaccination for bovine respiratory viruses on cattle health and subsequent bovine respiratory disease morbidity has been widely studied yet questions remain regarding the impact of these vaccines on host response and gene expression. Six randomly selected calves were vaccinated twice preweaning (T1 and T3) with a modified live vaccine for respiratory pathogens and 6 randomly selected calves were left unvaccinated. Whole blood samples were taken at first vaccination (T1), seven days later (T2), at revaccination and castration (T3), and at weaning (T4), and utilized for RNA isolation and sequencing. Serum from T3 and T4 was analyzed for antibodies to BRSV, BVDV1a, and BHV1. Sequenced RNA for all 48 samples was bioinformatically processed with a HISAT2/StringTie pipeline, utilizing reference guided assembly with the ARS-UCD1.2 bovine genome. Differentially expressed genes were identified through analyzing the impact of time across all calves, influence of vaccination across treatment groups at each timepoint, and the interaction of time and vaccination. Calves, regardless of vaccine administration, demonstrated an increase in gene expression over time related to specialized proresolving mediator production, lipid metabolism, and stimulation of immunoregulatory T-cells. Vaccination was associated with gene expression related to natural killer cell activity and helper T-cell differentiation, enriching for an upregulation in Th17-related gene expression, and downregulated genes involved in complement system activity and coagulation mechanisms. Type-1 interferon production was unaffected by the influence of vaccination nor time. To our knowledge, this is the first study to evaluate mechanisms of vaccination and development in healthy calves through RNA sequencing analysis.

KEYWORDS

beef calves, vaccination, preweaned calves, bovine respiratory disease, immunity, RNA sequencing (RNA-Seq), T-cell, inflammation

Introduction

Vaccination remains one of the most important tools for controlling bovine respiratory disease (BRD) in beef calves (1). Increasing adaptive immunity against known pathogens is the goal of vaccination; however, the presentation of antigens to the immune system elicits multiple cascading events within an animal as part of both innate and adaptive immunity. The most commonly measured indicators of adaptive immunity are serum antibody titers to specific pathogens of interest. Although useful, antibody titers have some limitations including multiple samples must be taken for accurate diagnosis of infection by endemic respiratory agents, and sufficient time must pass for the immune system to respond adequately (2, 3). One alternative to antibody titers is to evaluate differential gene expression to identify markers that indicate immune competency, immune responsiveness, and/or predict future immunity to those pathogens. However, knowledge gaps remain regarding the impact of vaccination on gene expression and how that gene expression may correlate with the development of adaptive immunity.

Although commercially available vaccines have been evaluated and approved by the USDA APHIS Center for Veterinary Biologics for purity, safety, potency, and efficacy, the requirements for efficacy studies are often quite different than the conditions under which the vaccine will be used in the field. While challenge studies can be very useful tools (4), they do not accurately model natural bovine respiratory disease and are often done with seronegative calves that have not been exposed to any pathogens or stressors. Strict protocols and timing of administration are also followed. In contrast, beef producers often use vaccines at different intervals from the label, in animals with a variety of backgrounds and nutritional, immune function, or passive transfer status, and often in the face of bacterial or viral exposure or other stressors (5). These differences can make it difficult to achieve the efficacy seen in the tightly controlled approval studies and raises the question whether vaccines, as commercially employed, are influencing rates of morbidity and performance in a consistent manner. To answer this question, the cattle industry needs additional research on these vaccines as they are employed in natural field conditions and the impact they have on cattle health, performance, and immune function.

Given this background, our objective was to explore differences in host gene expression in calves that were vaccinated preweaning with a modified live vaccine for respiratory pathogens or not *via* time-course transcriptomics, and to pair those data with antibody titers and health records. These data will support exploration of associations and generation of hypotheses regarding the immune response to vaccination that may influence future research and use of vaccines in preweaned beef calves.

Materials and methods

Animal use and study enrollment

All animal use and procedures were approved by the Mississippi State University Animal Care and Use Committee (IACUC protocol #19-169) and carried out in accordance with relevant IACUC and agency guidelines and regulations. This study was carried out in accordance with Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines (6).

Eighty-four bull calves were enrolled in a split plot design study to evaluate the impact of different management strategies on BRD morbidity, mortality, and performance (7). Animals were randomly assigned to whole plot (VAX or NOVAX) which were housed in 6 pastures during the cow-calf phase with no nose-to-nose contact. They were also randomly assigned to split plot level treatment of being directly transported to Texas for backgrounding after weaning (DIRECT) or sent to an auction market and then an order buyer facility for 3 days prior to transport to Texas for backgrounding (AUCTION); this event occurred after the timepoint T4, described below. All animals were visually assessed each day for signs of BRD and/or other disease by trained university employees and detailed health histories were kept on each calf. The observed signs of BRD were assigned a severity score of 0–4, adapted from the scoring system previously described by Holland et al. (8).

Calves were evaluated at four time points, described as T1, T2, T3, and T4. At T1, calves were vaccinated with 2 ml Pyramid 5 (Boehringer Ingelheim Animal Health) subcutaneously (VAX) or given 2 ml 0.9% saline subcutaneously (NOVAX) (median age = 107 days). Additionally, calves were tested *via* ear notch ELISA to evaluate PI status at T1; no PI positive calves were found. At T2, or 7 days post-vaccination (median age = 114 days), all calves were weighed and sampled. At T3 VAX calves were again administered (revaccinated with) 2 ml Pyramid 5 subcutaneously and NOVAX calves were given 2 ml 0.9% saline subcutaneously (median age = 183 days); all calves were castrated by knife with no analgesia on T3. All calves also received 5 ml of a multivalent clostridial bacterin-toxoid (Covexin 8, Merck Animal Health) subcutaneously at time point T1 and T3). All calves were handled so that no contact between vaccinated and non-vaccinated calves would occur. Calves were abruptly weaned at T4 (median age = 230 days) and entered the next phase of the study where they were kept in their original pastures in Mississippi for 3 days before being transported directly from Mississippi to Texas for backgrounding (DIRECT) or sent to an auction market where they stayed in a pen not in contact with other cattle for approximately 6 h, and then were transported for housing at an order buyer facility for 3 days prior to transport to Texas (AUCTION) Non-study calves from other sources were housed at the order buyer facility at the same time as the study calves, but they were not mixed with the study calves. In Texas (samples not evaluated in this study), calves were kept in one of 12 pens corresponding to their original random assignment to whole and split plot treatments ($n = 3$ pens of each pair of treatments). Whole blood was collected from all calves into Tempus RNA blood tubes (Applied Biosystems) and into serum tubes *via* jugular venipuncture immediately prior to first vaccination (T1), seven days post-vaccination (T2), immediately prior to vaccine booster administration and castration (T3), and at time of abrupt weaning (T4; 47 days post-booster). Overall, there were 7 days between T1 and T2, 70 days between T2 and T3, and 47 days between T3 and T4.

Twelve calves that remained clinically unaffected by BRD during the cow-calf and backgrounding phases of production were selected for RNA sequencing *via* stratified random sampling within the calves that remained healthy throughout the study within each backgrounding pen which resulted in 1 calf selected per backgrounding pen ($n = 3$ VAX/DIRECT, $n = 3$ NOVAX/DIRECT, $n = 3$ VAX/AUCTION, and $n = 3$ NOVAX/AUCTION). A total of 48 blood samples across the four time points were analyzed for whole blood transcriptomes. Metadata for all selected calves are found in Supplementary material 1.

Antibody titers

Serum collected at T3 and T4 was stored at −20°F before analysis at the University of Georgia's Athens Veterinary Diagnostic Laboratory. Serum neutralizing antibodies were assayed for bovine herpesvirus−1 (BHV-1), bovine viral diarrhea virus type 1a (BVDV1a), bovine respiratory syncytial virus (BRSV), and parainfluenza-3 virus (PI-3) per SOP # Ser013. Resulting titer levels for these antibodies are found in Supplementary material 1 and is limited to descriptive analysis only due to the small number of calves ($n = 12$).

Average daily gain

Differences in average daily gain between T1 and T4 were evaluated *via* generalized linear mixed effect models estimated *via* restricted pseudolikelihood with the Kenward-Rodgers adjustment for degrees of freedom in SAS 9.4. The model included vaccination status as a fixed effect and a random intercept for backgrounding pastures. Differences in least square means are reported and a cutoff of $p \leq 0.05$ was used to determine significance.

Next-generation RNA sequencing and bioinformatic data processing

Total RNA isolation, quality control, sequencing library preparation, and sequencing was performed by the Texas A&M University Institute for Genome Sciences and Society (TIGSS; College Station, TX, USA). Total RNA was isolated with Tempus Spin RNA Isolation Kit (Applied Biosystems), based on manufacturer's instructions. Total RNA from each sample was analyzed for RNA concentration and integrity with a Qubit 2.0 Fluorometer (ThermoFisher) and an Agilent 2,200 Bioanalyzer (Agilent), respectively; all RNA samples were of high quality (RIN: 7.8–9.5; mean = 8.8, s.d. = 0.3) and concentrations (ng/µL: 84.1–380.0; mean = 222.4, s.d. = 71.4). Library preparation for mRNA was performed with the TruSeq Stranded mRNA Library Prep Kit (Illumina), following manufacturer's instruction. Paired-end sequencing for 150 base pair read fragments was performed on an Illumina NovaSeq 6000 analyzer (v1.7+; S4 reagent kit v1.5) in one flow cell lane.

Quality assessment of reads was performed with FastQC v0.11.9[1] and MultiQC v1.12 (9), and read pair trimming for unambiguous base calls, adaptors, and retained minimum read length of 28 bases was performed with Trimmomatic v0.39 (10). Trimmed reads were mapped and indexed to the bovine reference genome assembly ARS-UCD1.2 with HISAT2 v2.2.1 (11). Sequence Alignment/Map (SAM) files were converted to Binary Alignment Map (BAM) files, prior to transcript assembly, with Samtools v1.14 (12). Transcript assembly and gene-level expression estimation for differential expression analysis was performed with StringTie v2.1.7 (13), as described by Pertea et al. (14). All sequencing data produced in this study are available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO), under the accession number GSE205004.

Differential gene expression analysis

Gene-level count matrices were processed and analyzed in RStudio, using R v4.1.2. Samples were classified by vaccination group and time point, where raw gene counts were processed and filtered by procedures described by Chen et al. (15). Any gene with a minimum total count above 100 and a count-per-million (CPM) of 0.2 in at least twelve samples was retained for further analysis. Post filtering, the complete dataset was considered non-sparse, and therefore normalized for differential expression analysis with the trimmed mean of M-values method (TMM) (16). Tagwise dispersion estimates of gene counts were supplied into the Bioconductor package glmmSeq v0.1.0[2] for negative binomial mixed effect modeling of gene counts. The following linear mixed-effect model was fitted to account for time points and vaccination group as

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] https://github.com/KatrionaGoldmann/glmmSeq

fixed effects, and housing (pasture) and individual ID as random effects:

$$Model: \sim Timepoint * Vaccine * Timepoint : Vaccine \\ + (1|Pasture) + (1|ID)$$

Model adaptation allowed for the assessment of differentially expressed genes (DEGs) across timepoints, vaccine groups, and the interactions between timepoints and vaccine group, where *p*-values were adjusted for false discovery rates (FDR) with the Benjamini-Hochberg method; genes were considered significantly expressed with an FDR ≤ 0.05. Pairwise comparisons for DEGs between each vaccination groups at every time point and within each vaccination group across each time point was performed with edgeR v3.36.0 (15, 17), fitting genes under generalized linear model (GLM) framework and employing quasi-likelihood F-tests (QLF); pairwise gene comparisons were considered significant with an FDR ≤ 0.10.

Dimensional reduction and unsupervised clustering analyses

Heatmap, principal component, and clustering analyses were performed with all filtered and log2 count-per-million (log2CPM) values of TMM-normalized gene counts between all 48 samples. Heatmap and exploratory clustering analysis of samples, with respect to vaccination, time points, and individual IDs, were performed with the Bioconductor package pheatmap v1.0.12,[3] utilizing Canberra distances and Pearson correlation coefficients for unsupervised hierarchical clustering of samples and DEGs, respectively. Specifically, z-scores were calculated and utilized for heatmap analysis from log2CPM values of normalized (TMM) expression values. Gene expression was grouped into 48 distinct clusters with the k-means algorithm embedded within pheatmap; the number of clusters was determined from the Elbow method. High dimensional data exploration and reduction *via* principal component analysis (PCA) was conducted with the Bioconductor package PCAtools v2.0.0,[4] utilizing a correlation matrix; normalized gene counts were processed through mean centering and variance scaling. A scree plot was generated to determine the number of principal components (PCs) to retain for analysis, utilizing Elbow and Horn's parallel analysis methods (18). A Spearman's rank correlation matrix of retained PCs was constructed with metadata components from all samples, which included individual identification (ID), birthweight, age of animal for each sample (Age), housing pen at Mississippi (Pasture), vaccination group (Vaccine), sampling time point for each sample (Timepoint), and the slope of weight gain over time starting at birth (i.e., growth rate; GR); correlations were considered significant with an FDR $\leq$ 0.10. To determine genes which were driving the variation seen among each significantly correlated PC, a loadings plot was generated with the top/bottom 2% retained variables across each component loading range. A PCA biplot was constructed from the PCs with significant correlation to vaccination groups; data ellipses were calculated from multivariate t-distributions and encompassed 80% confidence levels of expressional t-distribution across each time point.

Functional enrichment analyses of DEGs

Differentially expressed genes were analyzed for functional enrichment of gene ontology (GO) terms, Reactome pathways, and KEGG pathways with KOBAS-i (19) (accessed May 2, 2022), utilizing hypergeometric testing and Benjamini-Hockberg adjusted *p*-values (FDR $\leq$ 0.05). Functional enrichment of DEGs were analyzed in three separate analyses: (1) DEGs shared between time points in both glmmSeq and QLF testing of vaccinated and non-vaccinated calves (i.e., shared genes between glmmSeq–timepoints, QLF Vax T1vsT2, and QLF Novax T1vsT2), (2) DEGs identified between vaccination groups across each time point by both glmmSeq–vaccination and QLF testing (i.e., glmmSeq–Vaccine and Vax vs. Novax at T1), removing DEGs identified by method #1**,** and (3) DEGs solely identified in glmmSeq analysis of Timepoint: Vaccine interactions; this approach allowed for the independent assessment of functional enrichment influenced by calf development (i.e., time) and vaccine administration. Enriched GO terms and pathways were evaluated for directionality (increased or decreased) based on log2 fold changes of associated DEGs. Clustering and visualization of enriched KEGG terms was performed with the embedded enrichment visualization tool within KOBAS-i, utilizing edge (correlation) thresholds of 0.40 and top *n* clusters set to 8; more information regarding the embedded enrichment visualization tool framework is provided by Bu et al. (19).

Results

Antibody titers and average daily gain

Comparison of antibody titers indicated calves were likely naturally infected with BRSV and PI-3, because antibody titers to these agents increased between T3 and T4 in both vaccinated and non-vaccinated calves. Given the small number of calves evaluated, this somewhat clouds our ability to detect the effect of vaccination using serology of samples collected at only two timepoints. However, vaccinated calves appeared to respond well to MLV vaccination as indicated by the BVDV1a titer response (Supplementary material 1). Average daily gain between T1 and T4 was not significantly different ($p = 0.31$) between the

[3] https://CRAN.R-project.org/package=pheatmap

[4] https://github.com/kevinblighe/PCAtools

VAX (model-adjusted least square mean 1.89 lbs) and NOVAX (model-adjusted least square mean 2.13 lbs) groups.

Differential gene expression patterns and enriched biological mechanisms

Read mapping and alignment of the 48 transcriptomes to the ARS-UCD1.2 bovine reference genome resulted in an overall mapping rate average of 95.50% (s.d. = 0.96%). In total, gene-level alignment resulted in a total of 33,310 unique features, with a median library size of 41,089,614 (s.d. = 4,111,721) (Supplementary Figure 1). Pre-processing and filtering of low expression values resulted in a total of 17,371 genes used for downstream analyses (Supplementary Table 2). Analysis of genes from glmmSeq resulted in 1213, 435, and 85 DEGs when evaluating time, vaccination, and the interaction of time and vaccination, respectively (Supplementary Table 3). Comparative analyses for DEGs between vaccination groups and time points was conducted with edgeR GLM-QLF testing. Analysis of the NOVAX group over time yielded a total of 3,271 DEGs across six comparisons (Supplementary material 4). Analysis of the VAX group over time yielded a total of 4,085 DEGs across six comparisons (Supplementary material 5). Analysis of each time point between the VAX and NOVAX groups yielded a total of 861 DEGs across four comparisons (Supplementary material 6). Visualization of the number and directionality of DEGs identified from GLM-QLF testing and overlapping of DEGs from glmmSeq and edgeR QLF analyses, are found in Figure 1.

Heatmap and unsupervised clustering analysis, seen in Figure 2, demonstrated that the majority of calves (n = 7) were highly similar in global gene expression prior to vaccine administration (T1; right side). Time of sampling (Timepoint) emerged as a considerable factor in determining distinction between groups (Vaccine) and individual calves (ID), as the majority of samples on the left side of the heatmap (i.e., furthest from the T1 samples) were at time of vaccine boostering (T3) and weaning (T4). Several individuals (J015, J022, J027, J053, J109, J113, J124) demonstrated high self-similarity in global gene expression between time points.

Multidimensionality analysis and visualization of global gene expression patterns *via* PCA is found in Figure 3. Utilizing both the elbow method and Horn's parallel analysis, a total of 14 principal components (PCs) were determining as optimal for demonstrating explained variation across the 48 transcriptomes; the first 14 PCs retained 70.15% of the variance within the data (Figure 3A). Pairwise plotting of selective PCs (Figures 3B,C) was performed with those PCs which demonstrated significant correlations with timepoints and/or vaccination status (Figure 3D). The first PC, accounting for 14.20% of the total explained variance, was positively correlated with Age (r = 0.34, FDR < 0.10), Vaccine (r = 0.34, FDR < 0.10), and Timepoint (r = 0.44, FDR < 0.05). Two PCs, PC3 and PC4, accounting for 7.67 and 5.94% of total explained variance, respectively, demonstrated significant correlations with Timepoint but not Vaccine; PC3 demonstrated negative correlation with Timepoint (r = −0.38, FDR < 0.10) and ID (r = −0.39, FDR < 0.10) and PC4 demonstrated positive correlation with Timepoint (r = 0.38, FDR < 0.10) and Age (r = 0.47, FDR < 0.05), confounded by ID (r = −0.38, FDR < 0.10). Accounting for 5.74% of total explained variance, PC5 possessed significant negative correlation with Timepoint (r = −0.32, FDR < 0.10). While confounded by Pasture (r = −0.54, FDR < 0.01), PC10, accounting for 2.72% of total explained variance, possessed significant positive correlation with Vaccine (r = 0.36, FDR < 0.10). Notably, the strongest correlation found within this analysis was between PC11, accounting for 2.33% of total explained variance, and GR (r = 0.58, FDR < 0.01). The resulting pairwise plotting of PCs 1, 3, 4, 5, and 10 demonstrated relative overlapping of all samples at T1, with increasing dissimilarity of samples over time (Figure 3B). A biplot with statistical ellipses (multivariate C.I. = 80.00%) of the two PCs with significant correlation with Vaccine (PC1 and PC10) demonstrated high dissimilarity between timepoints T1 and T3, with relative high overlap of timepoints T2 and T4, with T3 variation driven by vaccinated calves J009, J022, J023, and J113 (Figure 3C). Genes driving the variation among each PC possessing significant metadata correlations are found in Figure 3E. Specifically, genes influencing variation within PC1 and PC10 (i.e., correlated PCs with Vaccination) include *AP5M1*, *CLOCK*, *EIF3K*, *HDAC3*, *MKLN1*, *MYNN*, *OCIAD1*, *PHIP*, *RACK1*, *RBM12B*, *RBM26*, *RPL37A*, *SNX17*, *STK16*, *TMEM208*, *TRAPPC1*, *UBXN7*, and *ZDHHC17* in PC1 and *KIR3DL1*, *LOC112447728*, and *LOC786987* in PC10, respectively. Those PCs having significant correlation with Timepoint, and not Vaccine (PC3, PC4, and PC5), possessed variance-driving genes which overlapped with glmmSeq–timepoint findings; *BATF*, *EXTL2*, *PRDX2*, *RNF122*, *TIAM1*, and *TMCC3* were identified in both PC3-5 loadings plots and glmmSeq–timepoint analysis.

Analysis of GO terms, KEGG pathways, and Reactome pathways of genes identified between glmmSeq–timepoints and edgeR QLF testing within both vaccination groups across time allowed for the assessment of enriched processes and pathways at three specific timepoint comparisons: (1) T1 vs. T3, (2) T1 vs. T4, and (3) T2 vs. T4 (Supplementary material 7). Shared DEGs from T1 vs. T3 comparisons enriched for 88 GO terms and 72 functional pathways. These GO terms were related to zinc ion binding, cytokine-mediated signaling, specifically interleukin-12, gene expression regulation, regulation to inflammatory response, including negative regulation of I-kappaB kinase/NF-kappaB signaling, and fatty acid metabolism and biosynthesis. Enriched pathways included the immune system (both innate and acquired immunity) retrograde endocannabinoid signaling,

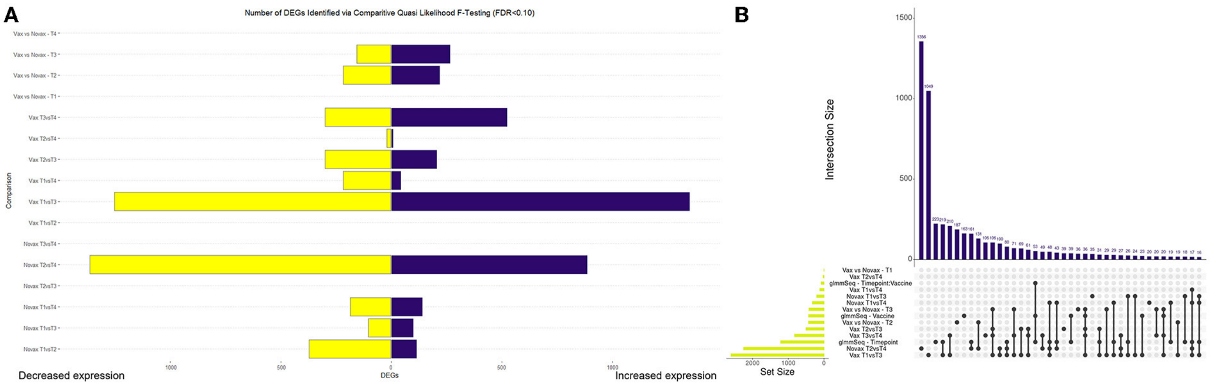


FIGURE 1
Visualization of differentially expressed genes (DEGs) identified through edgeR Quasi Likelihood F-testing and glmmSeq analyses. **(A)** Bar graph depicting the number and directionality of DEGs found in each edgeR pairwise test. Directionality is based on the first testing group within each pairwise test. For example, Vax T3vsT4 depicts 524 DEGs upregulated and 302 DEGs downregulated at T3 when compared to T4. **(B)** Upset plot demonstrating the number of DEGs overlapped between all differential expression analyses. Novax T2vsT4 possessed the most (1356) unique DEGs of any analysis, while Vax T1vsT3 and glmmSeq–Timepoint possessed the highest number of genes identified in multiple analyses (219).

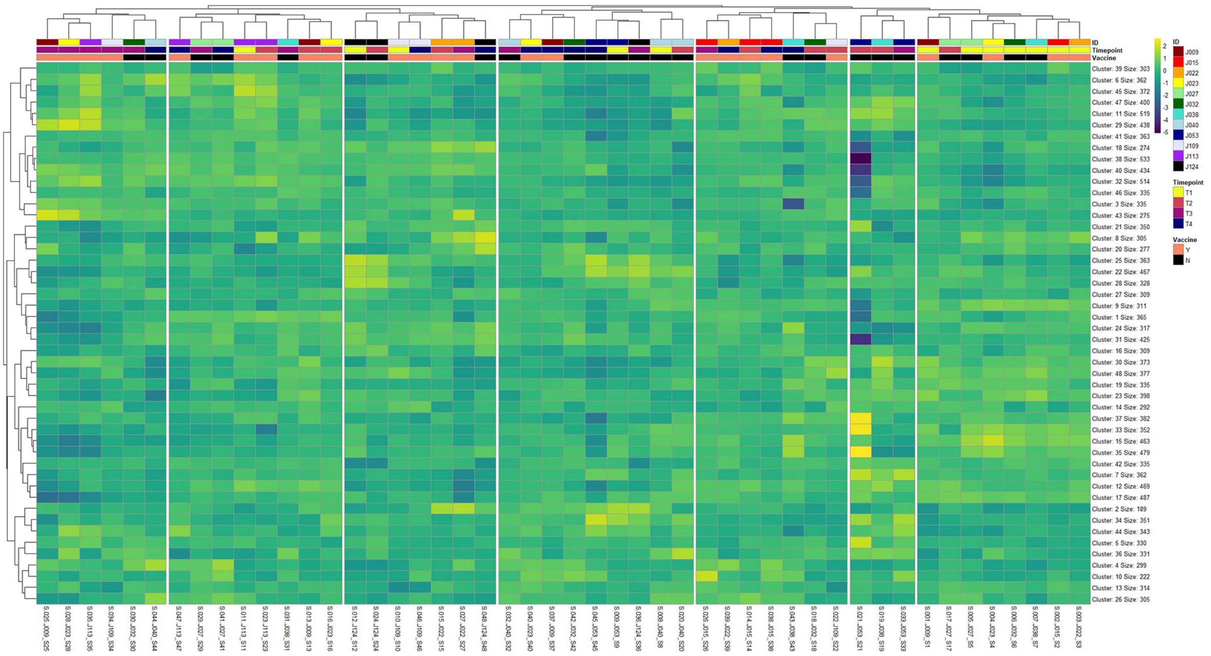


FIGURE 2
Heatmap and unsupervised hierarchical clustering analysis of global gene expression patterns across all 48 sample libraries ($n = 17{,}371$) following optimal k-means clustering of genes ($k = 48$). Gene clusters were labeled by clustering order (Cluster) and the total number of genes embedded within each cluster (Size). Sample libraries were labeled top-to-bottom with individual identification (ID), time point for each sample (Timepoint; T1, T2, T3, and T4), and vaccination group (Vaccine; Yes or No).

interleukin-4/13 signaling, glucose metabolism, glucagon signaling, TP53 expressional and degradation regulation, and the biosynthesis of specialized proresolving mediators (SPMs), including SPMs derived from both docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA). These GO terms and pathways were primarily enriched by the following DEGs: *ADAMTS12*, *ALOX15*, *ALOX5*, *CFL1*, *CPT1A*, *FBP1*, *FSCN1*, *IL5RA*, *LOC100297044* (*CCL14*), *LOC615278* (*TRIM39*), *LOC789732* (*CD300C*), *MIF*, *OTUD7B*, *PEG10*, *PIKFYVE*, *PLP2*, *POLR2L*, *PPP2R1A*, *PRKCG*, *PYGM*, *TK1*, and *TP53*.

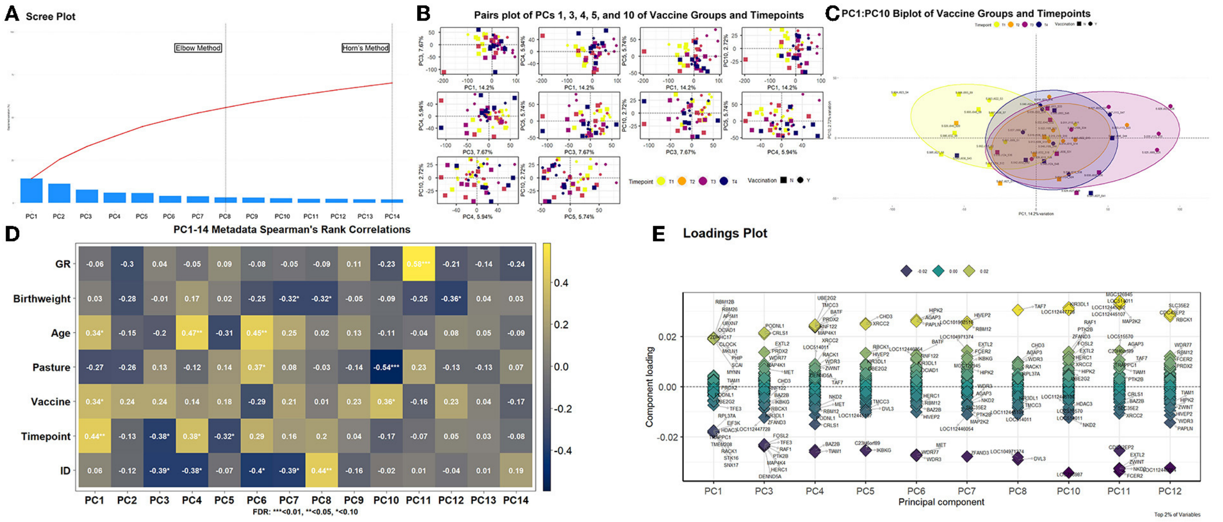


FIGURE 3

Principal component analysis of global gene expression patterns for all samples. **(A)** Scree plot[[Inline Image]] analysis depicting the maximum number of components to retain. Horn's parallel analysis method was ultimately utilized to retain the first 14 principal components (PCs), which explained 70.15% of total variance across the dataset. **(B)** Multiple biplot analysis (pairs plot) of PCs possessing significant correlation with timepoint and/or vaccination. Each point (vector) represents a PC score of an individual sample, in which the further from plot-center the point is, the more variation that sample contributes to the total variation. The colors yellow, orange, violet, and blue represent the timepoints T1, T2, T3, and T4, respectively; the shapes square or circle represent the vaccination status as no or yes, respectively. **(C)** Specific multivariate biplot analysis of PC1 and PC10, as influenced by timepoint and vaccination (see 3B color and shape coding). **(D)** Spearman's Rank correlation matrix heatmap of retained PCs and corresponding metadata components. Metadata components included the slope of weight gain over time (i.e., growth rate; GR), weight at birth (Birthweight), age at sampling (Age), pasture assignment (Pasture), vaccination status (Vaccine), time of sampling (Timepoint), and individual identification (ID). **(E)** Loading plot analysis with associated genes driving the variation explained by PCs with a significant correlation identified by 3D. Only the top 2% of genes based on component loading scores (i.e., most responsible for explained variation) were retained for each PC.

Shared DEGs from T1 vs. T4 comparisons enriched for 34 GO terms and 35 functional pathways. These GO terms were related to inflammatory response, cytokine-mediated signaling, magnesium ion binding, cellular response to oxidative stress, positive regulation of autophagy, T-cell co-stimulation, and actin/microtubule organization and development. Enriched pathways included the acquired immune system, interleukin signaling, cellular stress response, CD28 co-stimulation and signaling, and gap junction trafficking and regulation. These GO terms and pathways were primarily enriched by the following DEGs: *ALOX15*, *CD80*, *HMGA1*, *HSPB8*, *IL17REL*, *IL5RA*, *LOC100297044* (*CCL14*), *LOC533307* (*LRRK2*), *LOC789732* (*CD300LD*), *MAP3K8*, *NCF2*, *SLC7A11*, *TUBB*, *TUBB3*, and *ZC3H12A*. Shared DEGs from T2 vs. T4 comparisons enriched for 94 GO terms and 29 functional pathways. These GO terms were related to inflammatory and cytokine-mediated response, specifically including interleukin-17 receptor activity, MHC class I protein complex binding, response to mercury and magnesium ions, antigenic stimuli and macrophage differentiation, and fatty acid metabolism and biosynthesis. Enriched pathways included cellular metabolism involving fructose, mannose, pyruvate, and lipid metabolism, cytokine-cytokine receptor interaction, and the biosynthesis of specialized proresolving mediators (SPMs), including SPMs derived from both docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA). These GO terms and pathways were primarily enriched by the following DEGs: *ALOX15*, *CEBPE*, *DECR2*, *FBP1*, *IL17REL*, *IL5RA*, *LOC100297044* (*CCL14*), *LOC788694* (*KLRC1*), and *SLC7A11*. Visualization of the enriched KEGG pathway terms is found in Figure 4. Expressional trends of DEGs identified in all three timepoint comparisons between the two vaccination groups (*ALOX15*, *IL5RA*, *IL17REL*, *LOC100297044* (*CCL14*), and *SCL7A11*) are found in Figure 5.

A total of 435 genes were identified by glmmSeq-Vaccination to be differentially expressed (Supplementary material 3), with 109 unique DEGs identified by overlapping glmmSeq–Vaccination and GLM-QLF testing results, post-removal of DEGs identified in Timepoint evaluation (Supplementary material 8). Specifically, a total of one, 24, and 92 DEGs were identified between vaccination groups at timepoints T1, T2, and T3, respectively; no genes were found to be differentially expressed between vaccinated and non-vaccinated calves at T4 (Supplementary material 8). Only one DEG was identified at T1 (*HEXDC*; increased in Vaccinated) between vaccinated and non-vaccinated calves, therefore possessed no enriched GO terms nor pathways. Shared DEGs identified at T2 between vaccinated and non-vaccinated calves enriched for 139 GO terms and 61 functional

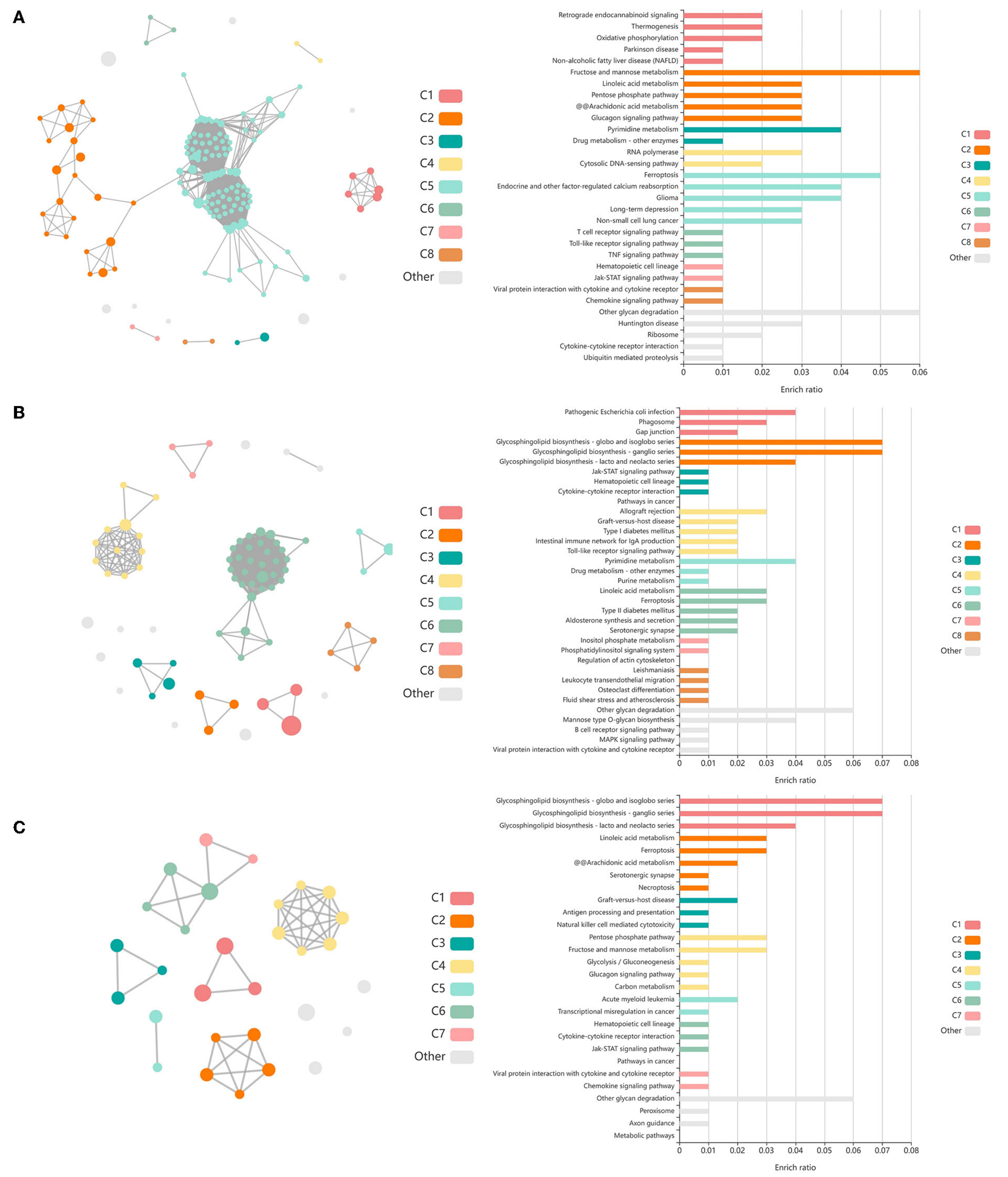


FIGURE 4
Clustering of enriched KEGG pathways by term identity from KOBAS-i analysis of DEGs influenced by time in both vaccination groups. Each node represents an enriched term, with color corresponding to the unique cluster based on term identity. Each edge (line between nodes) represents a significant correlation between pathway terms. Bar graphs represent the pathway terms found within each pathway (by color) and the level of enrichment (Enrich ratio). Gray nodes and bargraphed terms represent enriched pathways which did not associate within the clustering model. **(A)** KEGG pathways derived from T1 vs. T3 analysis clustered into eight unique clusters. **(B)** KEGG pathways derived from T1 vs. T4 analysis clustered into eight unique clusters. **(C)** KEGG pathways derived from T2 vs. T4 analysis clustered into seven unique clusters.

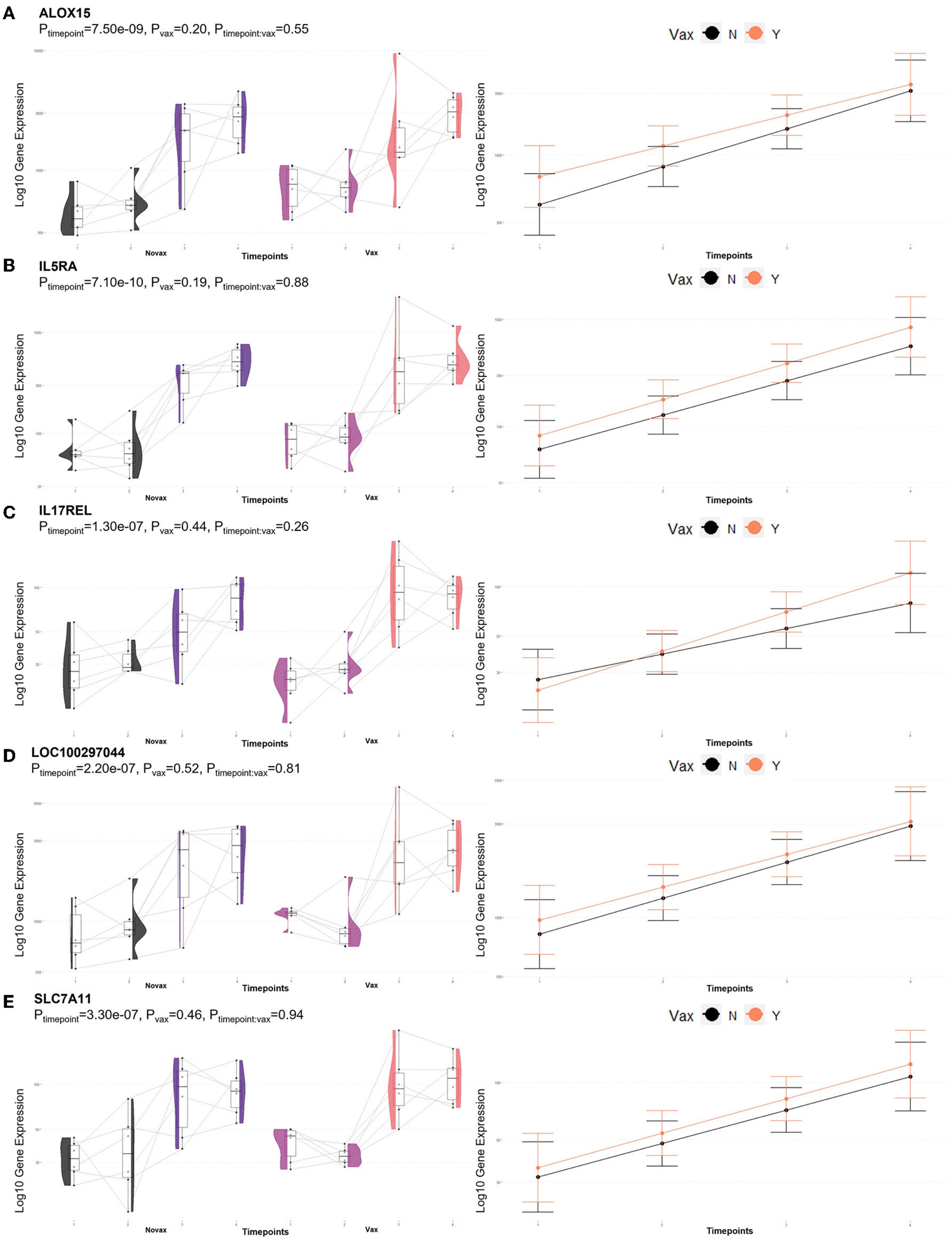


FIGURE 5
Gene pairplots and modeled expression trends of key DEGs found in timepoint analyses. Pairplots (left side) demonstrate the log10 normalized gene expression of each sample across all timepoints, overlapped with a violin plot (depicting numerical distributions by density). Box-and-whisker plots represent median expression values (black line), the first (lower) and third quartiles (boxplot limits), 1.5 times the interquartile ranges (whiskers), and outlier expression levels for each timepoint (points outside whiskers). Modeled expression trends (right side) depict the overall differences between groups over each timepoint. Points represent the mean log10 normalized expression value for each group within a timepoint, and bars represent the standard error of log10 normalized expression for each group; orange represents the vaccinated group and black represents the non-vaccinated group. These plots depict the relative expression and glmmSeq level of significance for **(A)** *ALOX15*, **(B)** *IL5RA*, **(C)** *IL17REL*, **(D)** *LOC100297044* (*CCL14*), and **(E)** *SLC7A11*.

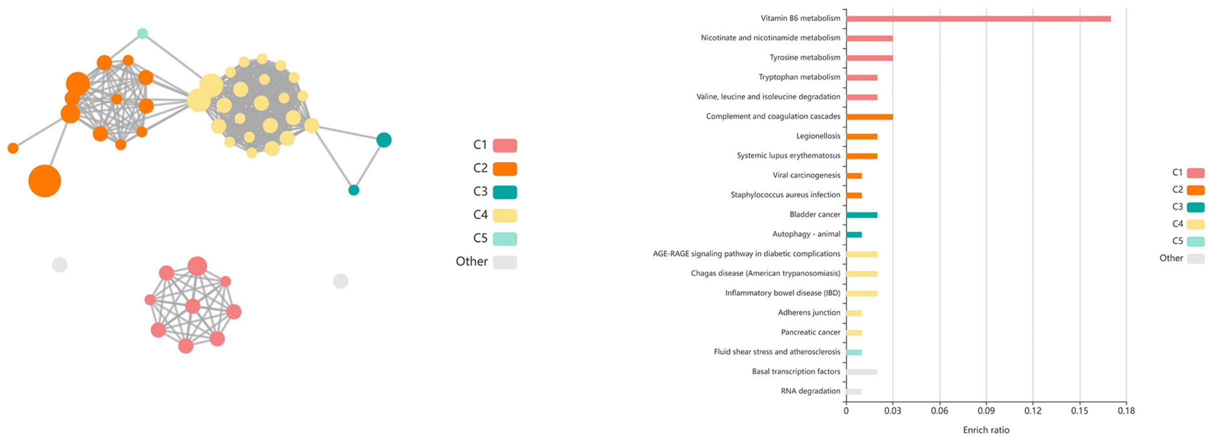


FIGURE 6
Clustering of enriched KEGG pathways by term identity from KOBAS-i analysis of DEGs identified between vaccinated and non-vaccinated calves at T2. Each node represents an enriched term, with color corresponding to the unique cluster based on term identity. Each edge (line between nodes) represents a significant correlation between pathway terms. Bar graphs represent the pathway terms found within each pathway (by color) and the level of enrichment (Enrich ratio). KEGG pathways identified between vaccine groups at T2 clustered into five unique clusters. Gray nodes and bar graphed terms represent enriched pathways which did not associate within the clustering model.

pathways. These GO terms were related to immune response and regulation (increased in Vaccinated), T-cell activation (increased in Vaccinated), metal ion binding (increased in Vaccinated), positive transcriptional regulation and protein processing (increased in Vaccinated), cellular proliferation and maintenance (increased in Vaccinated), complement activity (decreased in Vaccinated), and apoptotic clearance and phagocytosis (decreased in Vaccinated). Enriched pathways included the immune system and cytokine signaling, including interleukin-37 signaling (increased in Vaccinated), complement and coagulation cascades (decreased in Vaccinated), enhanced transcriptional activity, largely involving RNA polymerase II (increased in Vaccinated), and vitamin B6 metabolism (decreased in Vaccinated). These GO terms and pathways were primarily enriched by the following DEGs: *ARL4D*, *C3*, *CNOT4*, *GTF2A1*, *LOC785873* (*TRIM26*), *POU2F1*, *PUS10*, *SMAD3*, *THBD*, and *ZBTB41*. Visualization of the enriched KEGG pathways is found in Figure 6. Expressional trends of the aforementioned DEGs contributing to these GO terms and pathways are found in Figure 7.

Shared DEGs identified at T3 between vaccinated and non-vaccinated calves enriched for 71 GO terms and 25 functional pathways. These GO terms were related to neutrophil degranulation (increased in Vaccinated), antigen processing and presentation (increased in Vaccinated), ubiquitin protein binding and positive regulation (increased in Vaccinated), nuclear protein importing and response to protein folding (increased in Vaccinated), heat shock protein binding, specifically to Hsp70 and Hsp90 (increased in Vaccinated), T-cell activation (increased in Vaccinated), cellular response to interleukin-7 (increased in Vaccinated), and the positive regulation to ATPase activity (increased in Vaccinated). Enriched pathways included transcription activation (increased in Vaccinated), endocytosis and antigen processing and presentation (increased in Vaccinated), neutrophil degranulation (increased in Vaccinated), and cellular response to heat stress, including the regulation of HSF1-mediated heat shock response, Hsp90 chaperone cycle for steroid hormone receptors, and HSF1-dependent transactivation (all increased in Vaccinated). These GO terms and pathways were primarily enriched by the following DEGs: *AHSA2*, *BANP*, *C3*, *CACYBP*, *CCT2*, *DNAJA4*, *DNAJB1*, *DNAJB4*, *HIST1H3G*, *HSP90AB1*, *HSPA14*, *HSPA1A*, *HSPA4*, *HSPA6*, *HSPD1*, *HSPH1*, *KAT2A*, *MDM4*, *NUTF2*, *PTPRB*, *RAB11FIP3*, *RCHY1*, *SMAD3*, *STIP1*, *SYMPK*, *TOMM34*, *TRAF2*, *ZFAND2A*, *ZFP28*, and *ZNF473*. Visualization of the enriched KEGG pathway terms is found in Figure 8. Expressional trends of the DEGs primarily involved in immune mediated and heat shock response associated GO terms and pathways (*DNAJB1*, *DNAJB4*, *HSP90AB1*, *HSPA14*, *HSPA1A*, *HSPA4*, *HSPA6*, *HSPD1*, *HSPH1*, and *TRAF2*) are found in Figure 9.

A total of 85 DEGs were identified by glmmSeq when evaluating the interaction between Vaccination and Timepoints, which enriched for 13 GO terms and six functional pathways (Supplementary material 9). These GO terms were related to extracellular space, actin filament organization, cytoplasmic vesicles, copper ion binding, and natural killer cell activation and mediated cytotoxicity. Enriched pathways included small molecule transport, immunoregulatory interactions between lymphoid and non-lymphoid cells, plasma lipoprotein remodeling, natural killer cell mediated cytotoxicity, tyrosine

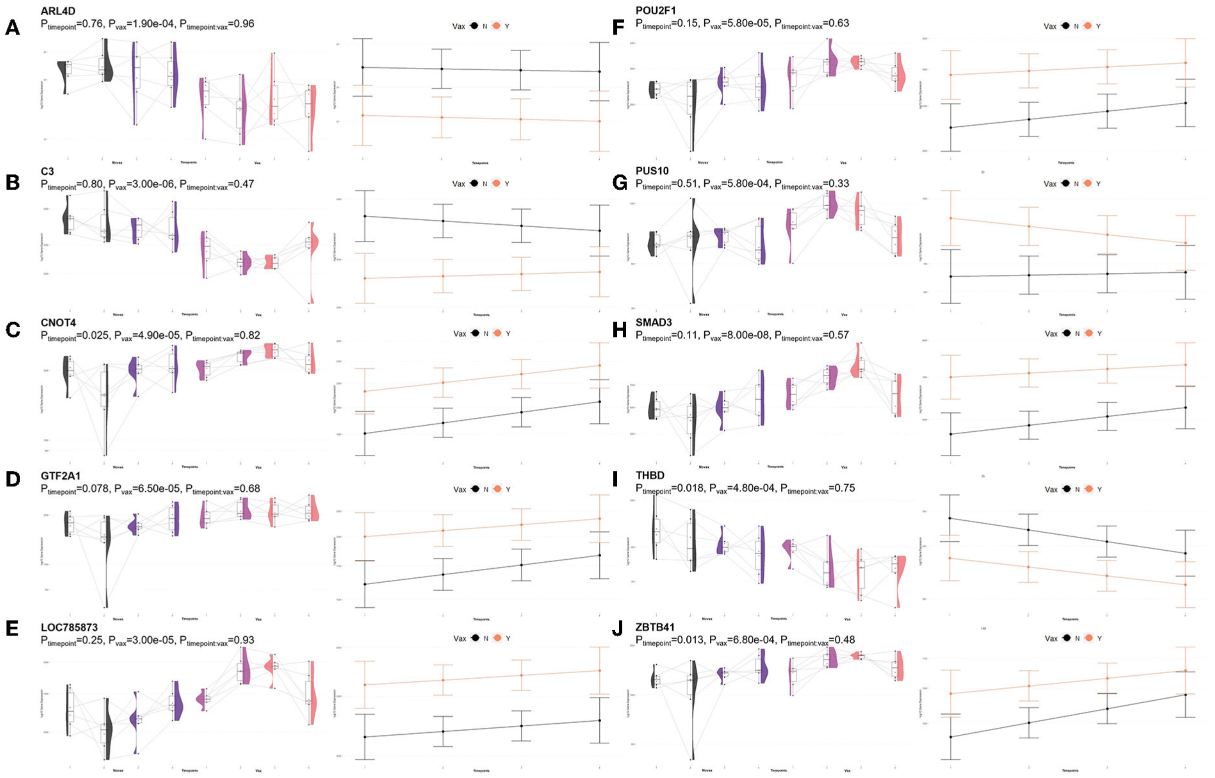


FIGURE 7
Gene pairplots and modeled expression trends of key DEGs found in timepoint analyses. Pairplots (left side) demonstrate the log10 normalized gene expression of each sample across all timepoints, overlapped with a violin plot (depicting numerical distributions by density). Box-and-whisker plots represent median expression values (black line), the first (lower) and third quartiles (boxplot limits), 1.5 times the interquartile ranges (whiskers), and outlier expression levels for each timepoint (points outside whiskers). Modeled expression trends (right side) depict the overall differences between groups over each timepoint. Points represent the mean log10 normalized expression value for each group within a timepoint, and bars represent the standard error of log10 normalized expression for each group; orange represents the vaccinated group and black represents the non-vaccinated group. These plots depict the relative expression and glmmSeq level of significance for **(A)** *ARL4D*, **(B)** *C3*, **(C)** *CNOT4*, **(D)** *GTF2A1*, **(E)** *LOC785873 (TRIM26)*, **(F)** *POU2F1*, **(G)** *PUS10*, **(H)** *SMAD3*, **(I)** *THBD*, and **(J)** *ZBTB41*.

metabolism, and DAP12 interactions. Visualization of the enriched KEGG pathway terms is found in Figure 10. Expressional trends of the DEGs primarily involved in immunoregulatory and natural killer cell associated GO terms and pathways (*AOC3*, *DCT*, *LOC100852061* (*KIR2DS2*), *LOC101905165* (*NKG2D*), *LOC112441504* (*ULBP3*), and *LOXL4*) is found in Figure 11.

Discussion

Use of modified live viral respiratory vaccines in beef cattle production systems

The use of modified live viral (MLV) vaccines in beef cattle backgrounding and feeding operations remains one of the leading practices in managing risk of BRD in cattle populations (1). Multiple recent reviews have evaluated the peer-reviewed literature regarding the use of various vaccines for respiratory pathogens in beef cattle (20–22). However, vaccination is not always helpful (23) and questions remain regarding which cattle are most likely to benefit from vaccination and which may not. Assessment of the transcriptome may reveal new pathways that will explain why vaccination appears to prevent disease in certain situations but not others. The significance of some of the differences in observed gene expression between VAX and NOVAX calves is not yet clear, but provides a foundation for future studies to determine how multiple components of the immune response change following vaccination. To our knowledge, there is no comparable data set available.

Although variable in terms of individual efficacy, several studies suggest that vaccinating herds of cattle with MLV vaccines reduces herd-level risk of BRD-associated morbidity and mortality and is associated with improved weight gain overtime (i.e., production) (24–27). Our study was limited to a small subset of calves that remained clinically healthy

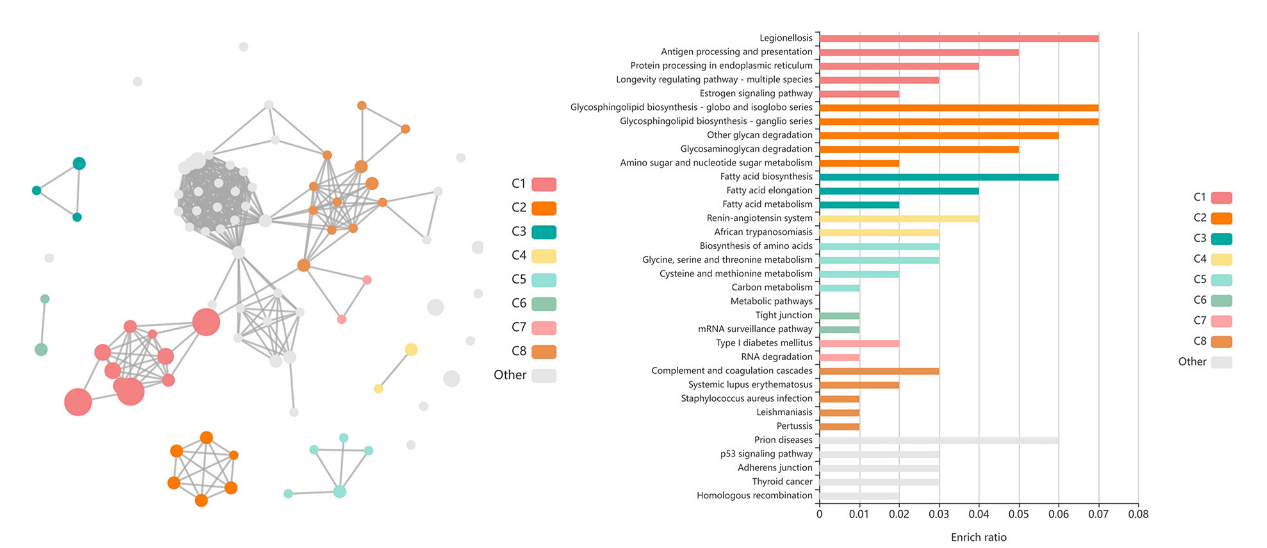


FIGURE 8
Clustering of enriched KEGG pathways by term identity from KOBAS-i analysis of DEGs identified between vaccinated and non-vaccinated calves at T3. Each node represents an enriched term, with color corresponding to the unique cluster based on term identity. Each edge (line between nodes) represents a significant correlation between pathway terms. Bar graphs represent the pathway terms found within each pathway (by color) and the level of enrichment (Enrich ratio). KEGG pathways identified between vaccine groups at T3 clustered into eight unique clusters. Gray nodes and bar graphed terms represent enriched pathways which did not associate within the clustering model.

which may have influenced the subsequent lack of difference in performance.

For years, responses to vaccination have been measured *via* serology (4), and occasionally, cell-mediated immune responses (20). Such studies usually describe only a small number of outcomes of a vast and diverse network of interactions that influence health vs. disease. While we attempted to assess serum neutralizing titer responses to the viruses we vaccinated against, the timing of sample collection was not optimized to find peak titer responses. Additionally, it is important to note that in this study the MLV was administered differently than the label directions indicated. The current label for Pyramid 5 (28) does not indicate a minimum age requirement or a specific interval or requirement for revaccination of calves. However, administering a booster vaccination, especially when the primary vaccination was given to animals under 6 months of age, is common industry practice and according to current knowledge would be helpful in initiating a protective immune response. Our serology results indicated that the calves responded to our vaccination strategy as expected but that there was likely a natural exposure to PI-3 and BRSV in the herd. The lack of differential gene expression at T4 between VAX and NOVAX calves is further supported by the titer data at T4. This may be due to the length of time between sampling points T3 and T4, but our data suggest both VAX and NOVAX individuals, across multiple pens, were exposed to a potentially non-virulent strain of PI-3 and BRSV sometime between T3 and T4. Furthermore, the similarities in gene expression between the two groups at T4 may be confounded due to this exposure and processing at T3. However, the results of this study demonstrated that the driver of immunological response and enhanced transcription over time, as influenced by vaccination, was the initial (first) administration.

Development of clinically healthy cattle is associated with increased specialized proresolving mediator expression, fatty acid and carbohydrate metabolism, and cytokine-mediated immunity

When evaluating the influence of time (i.e., physiological growth) on the gene expression of young calves, three connected mechanisms continually increased over time across all individuals: specialized proresolving mediator (SPM) biosynthesis, fatty acid and carbohydrate metabolism, and chemokine/cytokine mediated enhancement of acquired immunity. Specialized proresolving mediators consist of closely related classes of lipid mediators, derived from the lipoxygenation of arachidonic acid into LXA4 (i.e., lipoxins) (29) or from the metabolism of omega-3 and/or omega-6 essential polyunsaturated fatty acids (i.e., resolvins, protectins, and maresins) (30–32). Collectively, six molecules (*ALOX5*, *ALOX15*, *GPX4*, *HPGD*, *LTA4H*, and *PTGS2*) are directly involved in the biosynthesis of SPMs,[5] of which we identified

[5] https://reactome.org/PathwayBrowser/#/R-HSA-9018679&DTAB=MT

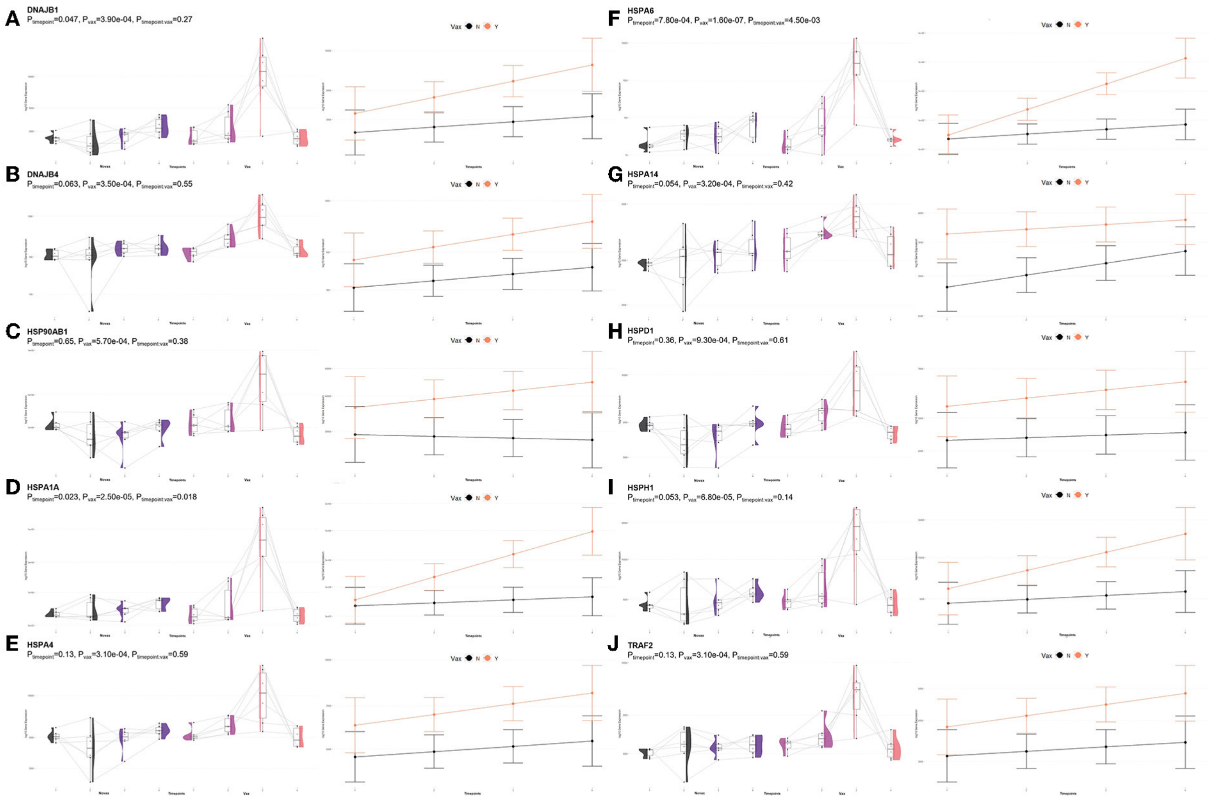


FIGURE 9
Gene pairplots and modeled expression trends of key DEGs found in timepoint analyses. Pairplots (left side) demonstrate the log10 normalized gene expression of each sample across all timepoints, overlapped with a violin plot (depicting numerical distributions by density). Box-and-whisker plots represent median expression values (black line), the first (lower) and third quartiles (boxplot limits), 1.5 times the interquartile ranges (whiskers), and outlier expression levels for each timepoint (points outside whiskers). Modeled expression trends (right side) depict the overall differences between groups over each timepoint. Points represent the mean log10 normalized expression value for each group within a timepoint, and bars represent the standard error of log10 normalized expression for each group; orange represents the vaccinated group and black represents the non-vaccinated group. These plots depict the relative expression and glmmSeq level of significance for **(A)** *DNAJB1*, **(B)** *DNAJB4*, **(C)** *HSP90AB1*, **(D)** *HSPA1A*, **(E)** *HSPA4*, **(F)** *HSPA6*, **(G)** *HSPA14*, **(H)** *HSPD1*, **(I)** *HSPH1*, and **(J)** *TRAF2*.

three to be differentially increased in all calves over time (*ALOX5*, *ALOX15*, and *HPGD*); notably, *HPGD* was identified as a differentially expressed in glmmSeq – timepoint, NOVAX T1vT4, VAX T1vT3, and VAX T2vT3. These lipid molecules are profound regulators of both acute and chronic inflammation and are critical in promoting cellular clearance and tissue remodeling in response to respiratory disease (33–36). Recent evidence suggests that, in addition to their ability to resolve inflammatory responses and tissue damage, SPMs are effective modulators of the adaptive immune response, capable of regulating Th1/Th17 differentiation and promoting regulatory T-cell differentiation *via* a non-cytopathic regulatory mechanism (37). Crucially, SPMs are shown to not have an effect on Th2-driven immunity, but enhance antigen presenting cell, specifically dendritic cell, development and functionality (37–39); this aligns with our findings indicating a gradual increase in gene expression related to SPM production and immunoregulatory T-cells. This is additionally supported by the enrichment of CD28 co-stimulation and signaling, and the enhancement of *CTLA4* and *CD80* with associated cytokine production (*IL5RA*, *IL17REL*) over time (40–44). Furthermore, several of these specific genes, namely *ALOX15*, *LOC100297044* (CCL14), *HPGD*, and *IL5RA*, have been identified as DEGs increased in expression at facility arrival in cattle that remain clinical healthy within high-risk populations, compared to cattle that develop BRD (45–49). Collectively, this may represent immunological development and mechanisms of immunocompetence which can serve a protective role against BRD-induced inflammation when calves are placed in post-weaned feeding systems.

Vaccination induces a controlled inflammatory response linked with Th17/natural killer cell activity

Evaluation of host expression influenced by vaccination, excluding genes and mechanisms affected solely by time,

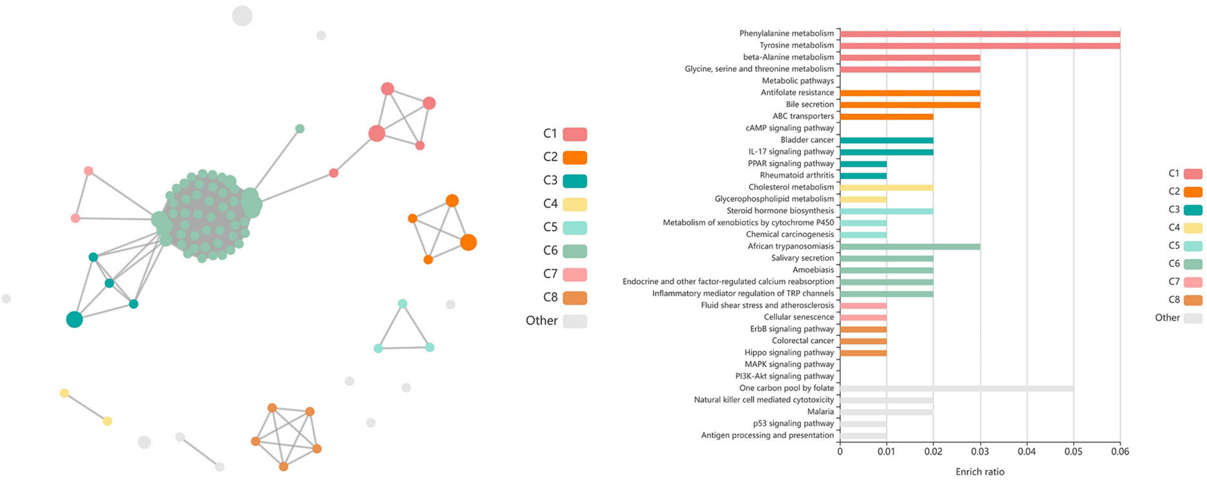


FIGURE 10
Clustering of enriched KEGG pathways by term identity from KOBAS-i analysis of DEGs identified from glmmSeq evaluation of the interaction between vaccination and time. Each node represents an enriched term, with color corresponding to the unique cluster based on term identity. Each edge (line between nodes) represents a significant correlation between pathway terms. Bar graphs represent the pathway terms found within each pathway (by color) and the level of enrichment (Enrich ratio). KEGG pathways identified through the interaction of vaccination and time clustered into eight unique clusters. Gray nodes and bar graphed terms represent enriched pathways which did not associate within the clustering model.

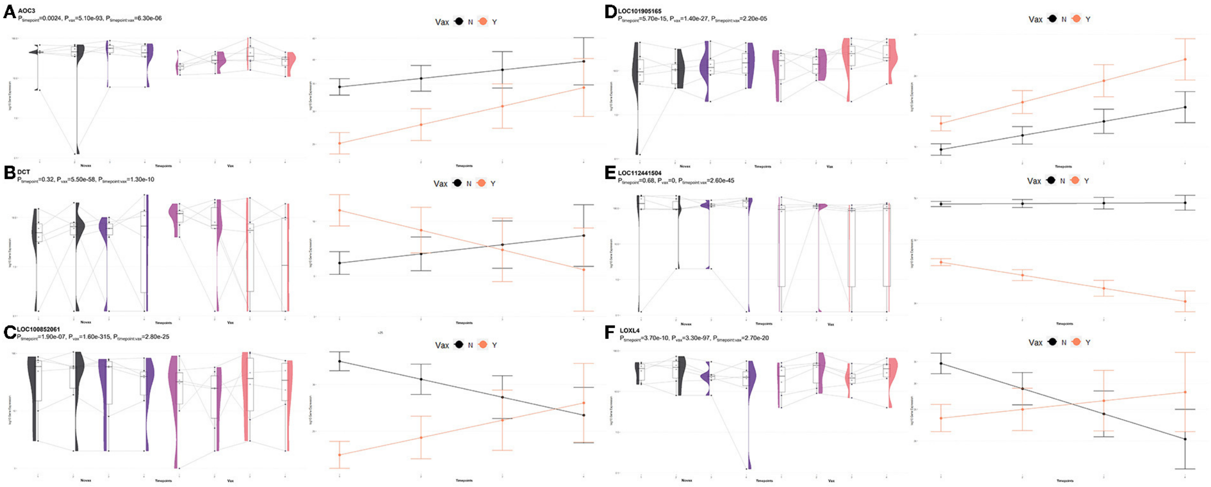


FIGURE 11
Gene pairplots and modeled expression trends of key DEGs found in timepoint analyses. Pairplots (left side) demonstrate the log10 normalized gene expression of each sample across all timepoints, overlapped with a violin plot (depicting numerical distributions by density). Box-and-whisker plots represent median expression values (black line), the first (lower) and third quartiles (boxplot limits), 1.5 times the interquartile ranges (whiskers), and outlier expression levels for each timepoint (points outside whiskers). Modeled expression trends (right side) depict the overall differences between groups over each timepoint. Points represent the mean log10 normalized expression value for each group within a timepoint, and bars represent the standard error of log10 normalized expression for each group; orange represents the vaccinated group and black represents the non-vaccinated group. These plots depict the relative expression and glmmSeq level of significance for **(A)** *AOC3*, **(B)** *DCT*, **(C)**, *LOC100852061* (*KIR2DS2*), **(D)** *LOC101905165* (*NKG2D*), **(E)** *LOC112441504* (*ULBP3*), and **(F)** *LOXL4*.

demonstrated an increase in mechanisms associated with antigen presentation, metal ion binding, molecular chaperone activity, and lymphoid cell activity, and a decrease in mechanisms associated with complement and apoptotic debris clearance. First, through PCA of global expression trends, we discovered genes driving variation in PCs with significant correlation with vaccination. Specifically, we identified the genes *CLOCK*, *HDAC3*, *KIR3DL1*, *RACK1*, and *SNX17* to be key drivers of differences associated with vaccination, which are involved in regulating the activity and differentiation of

T-cells and natural killer cells. *CLOCK*, in conjunction with *BMAL1*, is a circadian timekeeping protein which interacts with transcriptional regulators, which in turn upregulate genes such as *HDAC3*; this transcriptional network is responsible for the development and differentiation of Th17 cells (50–53). *KIR3DL1*, an immunoglobulin-like receptor expressed by natural killer cells and T-cells (54), is shown to be involved in inhibiting interferon-?? secretion and may block the progression of chronic inflammation, seen in research involving ankylosing spondylitis and reactive arthritis (55, 56). *RACK1*, which acts as both an intracellular protein receptor for protein kinase C and as a core ribosomal protein of the 40S subunit, is a key component of T-cell activation and proliferation (57, 58) and loss of *RACK1* has been shown to increase T-cell apoptosis (59). *SNX17* localizes with T-cell receptors and is responsible for preventing T-cell degradation into lysosomes and transporting T-cell receptors to the cell surface, aiding in cellular immune function (60). These findings provide initial evidence that vaccination in young calves influences mechanisms related to the enhanced differentiation and survival of T-cells, natural killer cell activity, and accompanying interleukin-17 response; this coincides with previous research demonstrating vaccination or exposure to viral components mediates a T-helper cell and natural killer cell response (61, 62), which may contribute to protective cell mediated and controlled inflammatory responses (63).

To further explore the influence of vaccination on these calves, we identified DEGs between vaccination groups at each time point, and those found from the interaction between vaccination and time. At T2 (7 days post vaccination), DEGs identified in calves which received a MLV vaccination enriched for two major immune-related mechanisms-the downregulation of complement and coagulation cascades (primarily driven by *C3*), and the upregulation of T-cell-mediated immunity. Complement, a well-organized and highly regulated system of the immune system, is a critical component of host immunity for killing or neutralizing pathogens and maintaining immunological homeostasis (64, 65). While the complement system features three distinct response pathways (classical, alternative, and mannose-binding lectin), all three lead to subsequent *C3* activation (66). Interestingly, the vaccinated group demonstrated a downregulation of *C3* transcription. While complement *C3* is critical for inducing a humoral and cell-mediated response to vaccination and viral infection (67–70), little published information exists relating to the timing and activity levels of induced complement cascades in cattle. Thus, it can be hypothesized that we failed to capture the initial immune responses associated with vaccination within the first few days and are identifying a late feedback mechanism involved in controlling prolonged complement activity. Additionally, research has demonstrated that the complement system appears to be more important for successful immunization in response to polysaccharide-containing vaccines compared to conjugated vaccines (71). Furthermore, we identified DEGs and enriched mechanisms related to CD4+ T-cell activity, primarily driven by *DAPK2*, *POU2F1*, *SMAD3*, and *LOC785873* (*TRIM26*). *DAPK2* promotes cellular recruitment to sites of inflammation (72) and is highly expressed in activated T-cells, serving a cellular regulatory role during germinal center formation (73). *POU2F1* is a required transcription factor for T-cell response to infection and the development of CD4+ memory T-cells (74–76). *SMAD3* transduces TGF-BR signaling and controls the development of regulatory T-cells and Th17 cells v*ia* signaling networks involving T-cell receptors, TGF-B, and interleukin-6 (77, 78). While not directly involved with T-cell activity, *TRIM26* is involved with modulating host antiviral defense and inducing an inflammatory immune response (79–81).

Evaluation of T3 (prior to booster administration; 77 days after initial vaccination) identified DEGs and enriched immunological mechanisms involved in neutrophil degranulation, antigen processing and interleukin-7 response, T-cell activation, transcriptional activity, and heat shock protein activity and binding. At this time point in vaccinated calves compared to non-vaccinated calves, we again observed an increase in the expression of *SMAD3* and *LOC785873* (*TRIM26*), two genes involved in T-cell development and inflammatory defense mechanisms, respectively, and a decrease in *THBD* and *C3*, involved in coagulation (82, 83) and complement activity, respectively. Surprisingly, these genes and associated mechanisms remained significantly enriched at both seven and 77 days post-vaccinated, indicating a possible immune-mediated mechanism or complex induced by MLV vaccination which persists longer than anticipated (>30 days). One unexpected finding at this timepoint was the rapid increase in heat shock protein gene expression in vaccinated calves. There is a great deal of research demonstrating the role of heat shock proteins in vaccination and host immunity/inflammation. Hsp70 enhances immunogenic antigen presentation cell functionality and T-cell proliferation (84). Both Hsp70 and Hsp90 proteins are shown to activate dendritic cells and direct naïve helper T-cell priming through designated interactions with antigen presenting cell surface receptors (85, 86) and stimulating inflammatory cytokine production *via* CD14-mediated chaperoning (87–89). *HSPD1* initiates interferon-beta production through interactions with interferon regulatory factor 3 (*IRF3*) (90) and is associated with both leukocyte and lymphocyte tissue infiltration (91). Furthermore, both Hsp70 and Hsp90 promote Th17 gene expression and proliferation and are involved in interleukin-17-mediated inflammation (92–94). This collectively indicates a stimulation of heat shock protein-mediated inflammation and helper T-cell, possibly Th17, promotion *via* modified live viral vaccination. However, this mechanism was only upregulated at T3. How long and where in time this mechanism becomes upregulated through viral vaccination could not be fully elucidated by this study and additional research is needed.

Our final differential expression evaluation was to determine genes influenced by the interaction of both time and vaccination. Largely, the DEGs identified through this analysis were determined to be involved with immunoregulatory functions *via* natural killer cells. These mechanisms were enriched by three key genes: *LOC101905165* (*NKGD2*), *LOC100852061* (*KIR2DS2*), and *LOC112441504* (*ULBP3*). *NKG2D* serves as a costimulatory transmembrane receptor on natural killer cells, enhancing T-cell receptor activity and subsequent cytotoxic and gamma-delta T-cell function (95–97). *KIR2DS2* is an activating receptor of natural killer cells, which binds to MHC class 1 and enhances natural killer cell-mediated cytotoxicity (98, 99). *ULBP3* is a cellular ligand of natural killer cells which binds to *NKG2D*, serving an immunostimulatory role (100–102). This indicates that the influence of both time and vaccination acts in influencing natural killer cell and cytotoxic responses in calves. Bassi et al. (103) demonstrated that cattle naturally infected with and displaying clinical signs of bovine papillomavirus possessed an increase in circulating natural killer cells and CD4+/CD8+ ratios, with a related elevation in interleukin-17 levels, when compared to cattle without clinical papillomatosis. Hamilton et al. (104) found that BCG vaccination in neonatal calves induces effector natural killer cells after interactions with dendritic cells, and stimulates their production of type-2 interferon production and interleukin-12.

Another key detail in this study is the lack of differential expression related to type-1 interferon production and response. Research has demonstrated that administration of recombinant and mRNA vaccines against viral pathogens can induce type-1 interferon production, enhancing T-cell response *via* heighted antigen presentation, and further promoting humoral immunity and vaccine-induced antibody production (105–108). Previous research in cattle has demonstrated that type-1 interferon production is strongly induced by viral challenge and is seen as an antiviral defense mechanism (109–112). While type-1 interferon production in human vaccination trials is well documented, it is relatively unknown if ungulates possess a similar immune response. We may have failed to recognize such a response due to the time between sampling points. Further studies assessing additional time points post-vaccination and booster, and focused assessment of peripheral immune cell types and responses, are warranted to better identify and understand the complex interaction of mechanisms related to successful immunization.

This study is, to our knowledge, the first of its kind to describe differential gene expression pathways in calves over the first 7 months of life, and in relation to a commonly used vaccination scheme, in a longitudinal fashion. Our findings indicate that vaccination induces a controlled inflammatory response associated with natural killer cell and, likely, Th17 cell promotion. This is most likely a normal process of antigen presentation and immunological memory within calves, but still constitutes an inflammatory-inducing process. It may be hypothesized that these induced mechanisms are not effective when calves are placed in high-risk settings, where stress and inflammation are occurring, compared to the low-risk system in which these calves were studied.

Data availability statement

The data presented in the study are deposited in the National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO), accession number GSE205004.

Ethics statement

The animal study was reviewed and approved by Mississippi State University Animal Care and Use Committee (IACUC protocol #19-169).

Author contributions

SC, AW, and BK conceived and designed the vaccination study funded by USDA. MS, AW, BK, and KH managed the live animal project and collected samples and metadata for analysis. MS and SC coordinated diagnostics, analyzed the data, and wrote the manuscript. SC, MS, AW, BK, and KH edited the manuscript and approved the final version. All authors agree to be accountable for the content of the work.

Funding

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2019-67015-29845 from the USDA National Institute of Food and Agriculture. Additionally, this project was supported in part by internal funds provided by Texas A&M AgriLife Research and Texas A&M University School of Veterinary Medicine and Biomedical Sciences.

Acknowledgments

The authors would like to acknowledge Libby Durst, Kalisha Yankey, Will Crosby, and Jane Parish for their assistance with managing the calves and collecting metadata at MSU. We thank Andrew Hillhouse and the staff at the Texas A&M Institute for Genome Sciences and Society for assistance with the RNA sequencing. We would also like to acknowledge Jill Fishburn and Binu Velayudhan at the University of Georgia Athens Diagnostic Lab for their assistance with the serum neutralizing antibody titers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture nor the internal supporters of this project.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.1010039/full#supplementary-material

References

1. Wilson BK, Richards CJ, Step DL, Krehbiel CR. Best management practices for newly weaned calves for improved health and well-being. *J Anim Sci.* (2017) 95:2170–82. doi: 10.2527/jas.2016.1006

2. Meeusen ENT, Walker J, Peters A, Pastoret P-P, Jungersen G. Current status of veterinary vaccines. *Clin Microbiol Rev.* (2007) 20:489–510. doi: 10.1128/CMR.00005-07

3. Savini G, Tittarelli M, Bonfini B, Zaghini M, Di Ventura M, Monaco F. Serological response in cattle and sheep following infection or vaccination with bluetongue virus. *Vet Ital.* (2004) 40:645–7.

4. Callan RJ, editor. Fundamental considerations in developing vaccination protocols. In *American Association of Bovine Practitioners Annual Conference.* Vancouver, Canada (2001).

5. Richeson JT, Hughes HD, Broadway PR, Carroll JA. Vaccination management of beef cattle: delayed vaccination and endotoxin stacking. *Vet Clin North Am Food Anim Pract.* (2019) 35:575–92. doi: 10.1016/j.cvfa.2019.07.003

6. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The Arrive guidelines 2.0: updated guidelines for reporting animal research. *PLoS Biol.* (2020) 18:e3000410. doi: 10.1371/journal.pbio.3000410

7. Capik SF, Amrine DE, White BJ, Larson RL, Woolums AR, Karisch BB, et al. P170 - Impact of management decisions on bovine respiratory disease morbidity and mortality risks. In: *Conference of Research Workers in Animal Diseases.* Chicago, IL (2019).

8. Holland BP, Step DL, Burciaga-Robles LO, Fulton RW, Confer AW, Rose TK, et al. Effectiveness of sorting calves with high risk of developing bovine respiratory disease on the basis of serum haptoglobin concentration at the time of arrival at a feedlot. *Am J Vet Res.* (2011) 72:1349–60. doi: 10.2460/ajvr.72.10.1349

9. Ewels P, Magnusson M, Lundin S, Käller M. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* (2016) 32:3047–8. doi: 10.1093/bioinformatics/btw354

10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

11. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat Biotechnol.* (2019) 37:907–15. doi: 10.1038/s41587-019-0201-4

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. *Bioinformatics.* (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352

13. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome Biol.* (2019) 20:278. doi: 10.1186/s13059-019-1910-1

14. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of Rna-Seq experiments with hisat, stringtie and ballgown. *Nat Protoc.* (2016) 11:1650–67. doi: 10.1038/nprot.2016.095

15. Chen Y, Lun AT, Smyth GK. From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Res.* (2016) 5:1438. doi: 10.12688/f1000research.8987.2

16. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of Rna-Seq data. *Genome Biol.* (2010) 11:R25. doi: 10.1186/gb-2010-11-3-r25

17. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor Rna-Seq experiments with respect to biological variation. *Nucleic Acids Res.* (2012) 40:4288–97. doi: 10.1093/nar/gks042

18. Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika.* (1965) 30:179–85. doi: 10.1007/BF02289447

19. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, et al. Kobas-I: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* (2021) 49:W317–w25. doi: 10.1093/nar/gkab447

20. Chamorro MF, Palomares RA. Bovine respiratory disease vaccination against viral pathogens: modified-live versus inactivated antigen vaccines, intranasal versus parenteral, what is the evidence? *Vet Clin North Am Food Anim Pract.* (2020) 36:461–72. doi: 10.1016/j.cvfa.2020.03.006

21. O'Connor AM, Hu D, Totton SC, Scott N, Winder CB, Wang B, et al. A systematic review and network meta-analysis of bacterial and viral vaccines, administered at or near arrival at the feedlot, for control of bovine respiratory disease in beef cattle. *Anim Health Res Rev.* (2019) 20:143–62. doi: 10.1017/S1466252319000288

22. Capik SF, Moberly HK, Larson RL. Systematic review of vaccine efficacy against mannheimia haemolytica, pasteurella multocida, and histophilus Somni in North American Cattle. *Bov. Pract.* (2021) 55:125–33. doi: 10.21423/bovine-vol55no2p125-133

23. Stokka GL. Prevention of respiratory disease in cow/calf operations. *Vet Clin Food Anim Pract.* (2010) 26:229–41. doi: 10.1016/j.cvfa.2010.04.002

24. Schumaher TF, Cooke RF, Brandão AP, Schubach KM, de Sousa OA, Bohnert DW, et al. Effects of vaccination timing against respiratory pathogens on performance, antibody response, and health in feedlot cattle1. *J Anim Sci.* (2018) 97:620–30. doi: 10.1093/jas/sky466

25. Step DL, Krehbiel CR, Burciaga-Robles LO, Holland BP, Fulton RW, Confer AW, et al. Comparison of single vaccination versus revaccination with a modified-live virus vaccine containing bovine herpesvirus-1, bovine viral diarrhea virus (Types 1a and 2a), parainfluenza type 3 virus, and bovine respiratory syncytial virus in the prevention of bovine respiratory disease in cattle. *J Am Vet Med Assoc.* (2009) 235:580–7. doi: 10.2460/javma.235.5.580

26. Palomares RA, Givens MD, Wright JC, Walz PH, Brock KV. Evaluation of the onset of protection induced by a modified-live virus vaccine in calves challenge inoculated with Type 1b bovine viral diarrhea virus. *Am J Vet Res.* (2012) 73:567–74. doi: 10.2460/ajvr.73.4.567

27. Kelling CL, Hunsaker BD, Steffen DJ, Topliff CL, Eskridge KM. Characterization of protection against systemic infection and disease from experimental bovine viral diarrhea virus type 2 infection by use of a modified-live noncytopathic type 1 vaccine in calves. *Am J Vet Res.* (2007) 68:788–96. doi: 10.2460/ajvr.68.7.788

28. LLC DS. *Pyramid 5 Boehringer Ingelheim Animal Health USA, Inc.* Available from: https://bayerall.cvpservice.com/product/view/1028246 (accessed July 22, 2022).

29. Levy BD, Clish CB, Schmidt B, Gronert K, Serhan CN. Lipid mediator class switching during acute inflammation: signals in resolution. *Nat Immunol.* (2001) 2:612–9. doi: 10.1038/89759

30. Serhan CN, Hong S, Gronert K, Colgan SP, Devchand PR, Mirick G, et al. Resolvins: a family of bioactive products of omega-3 fatty acid transformation circuits initiated by aspirin treatment that counter proinflammation signals. *J Exp Med.* (2002) 196:1025–37. doi: 10.1084/jem.20020760

31. Serhan CN, Yang R, Martinod K, Kasuga K, Pillai PS, Porter TF, et al. Maresins: novel macrophage mediators with potent antiinflammatory and proresolving actions. *J Exp Med.* (2009) 206:15–23. doi: 10.1084/jem.20081880

32. Serhan CN. Pro-resolving lipid mediators are leads for resolution physiology. *Nature.* (2014) 510:92–101. doi: 10.1038/nature13479

33. Nijmeh J, Levy BD. Lipid-derived mediators are pivotal to leukocyte and lung cell responses in sepsis and ards. *Cell Biochem Biophys.* (2021) 79:449–59. doi: 10.1007/s12013-021-01012-w

34. El Kebir D, József L, Pan W, Wang L, Petasis NA, Serhan CN, et al. 15-Epi-lipoxin a4 inhibits myeloperoxidase signaling and enhances resolution of acute lung injury. *Am J Respir Crit Care Med.* (2009) 180:311–9. doi: 10.1164/rccm.200810-1601OC

35. Levy BD, Kohli P, Gotlinger K, Haworth O, Hong S, Kazani S, et al. Protectin D1 is generated in asthma and dampens airway inflammation and hyperresponsiveness. *J Immunol.* (2007) 178:496–502. doi: 10.4049/jimmunol.178.1.496

36. Luo J, Zhang WY, Li H, Zhang PH, Tian C, Wu CH, et al. Pro-resolving mediator resolvin E1 restores alveolar fluid clearance in acute respiratory distress syndrome. *Shock.* (2022) 57:565–75. doi: 10.1097/SHK.0000000000001865

37. Chiurchiù V, Leuti A, Dalli J, Jacobsson A, Battistini L, Maccarrone M, et al. Proresolving lipid mediators resolvin D1, resolvin D2, and maresin 1 are critical in modulating T cell responses. *Sci Transl Med.* (2016) 8:353ra111. doi: 10.1126/scitranslmed.aaf7483

38. Basil MC, Levy BD. Specialized pro-resolving mediators: endogenous regulators of infection and inflammation. *Nat Rev Immunol.* (2016) 16:51–67. doi: 10.1038/nri.2015.4

39. Rothe T, Gruber F, Uderhardt S, Ipseiz N, Rössner S, Oskolkova O, et al. 12/15-lipoxygenase-mediated enzymatic lipid oxidation regulates dc maturation and function. *J Clin Invest.* (2015) 125:1944–54. doi: 10.1172/JCI78490

40. Walker LS, Sansom DM. The emerging role of Ctla4 as a cell-extrinsic regulator of T cell responses. *Nat Rev Immunol.* (2011) 11:852–63. doi: 10.1038/nri3108

41. Wu B, Jin M, Zhang Y, Wei T, Bai Z. Evolution of the Il17 receptor family in chordates: a new subfamily Il17rel. *Immunogenetics.* (2011) 63:835–45. doi: 10.1007/s00251-011-0554-4

42. Pätzold L, Stark A, Ritzmann F, Meier C, Tschernig T, Reichrath J, et al. Il-17c and Il-17re promote wound closure in a staphylococcus aureus-based murine wound infection model. *Microorganisms.* (2021) 9:1821. doi: 10.3390/microorganisms9091821

43. Hall BM, Hall RM, Tran GT, Robinson CM, Wilcox PL, Rakesh PK, et al. Interleukin-5 (Il-5) therapy prevents allograft rejection by promoting Cd4(+)Cd25(+) Ts2 regulatory cells that are antigen-specific and express Il-5 receptor. *Front Immunol.* (2021) 12:714838. doi: 10.3389/fimmu.2021.714838

44. Peach RJ, Bajorath J, Naemura J, Leytze G, Greene J, Aruffo A, et al. Both extracellular immunoglobin-like domains of Cd80 contain residues critical for binding T cell surface receptors Ctla-4 and Cd28. *J Biol Chem.* (1995) 270:21181–7. doi: 10.1074/jbc.270.36.21181

45. Sun HZ, Srithayakumar V, Jiminez J, Jin W, Hosseini A, Raszek M, et al. Longitudinal blood transcriptomic analysis to identify molecular regulatory patterns of bovine respiratory disease in beef cattle. *Genomics.* (2020) 112:3968–77. doi: 10.1016/j.ygeno.2020.07.014

46. Scott MA, Woolums AR, Swiderski CE, Perkins AD, Nanduri B, Smith DR, et al. Whole blood transcriptomic analysis of beef cattle at arrival identifies potential predictive molecules and mechanisms that indicate animals that naturally resist bovine respiratory disease. *PLoS ONE.* (2020) 15:e0227507. doi: 10.1371/journal.pone.0227507

47. Scott MA, Woolums AR, Swiderski CE, Perkins AD, Nanduri B, Smith DR, et al. Multipopulational transcriptome analysis of post-weaned beef cattle at arrival further validates candidate biomarkers for predicting clinical bovine respiratory disease. *Sci Rep.* (2021) 11:23877. doi: 10.1038/s41598-021-03355-z

48. Scott MA, Woolums AR, Swiderski CE, Thompson AC, Perkins AD, Nanduri B, et al. Use of ncounter mrna profiling to identify at-arrival gene expression patterns for predicting bovine respiratory disease in beef cattle. *BMC Vet Res.* (2022) 18:77. doi: 10.1186/s12917-022-03178-8

49. Scott MA, Woolums AR, Swiderski CE, Finley A, Perkins AD, Nanduri B, et al. Hematological and gene co-expression network analyses of high-risk beef cattle defines immunological mechanisms and biological complexes involved in bovine respiratory disease and weight gain. *bioRxiv.* (2022). doi: 10.1101/2022.02.16.480640 [preprint].

50. Man K, Loudon A, Chawla A. Immunity around the clock. *Science.* (2016) 354:999–1003. doi: 10.1126/science.aah4966

51. Yu X, Rollins D, Ruhn KA, Stubblefield JJ, Green CB, Kashiwada M, et al. Th17 cell differentiation is regulated by the circadian clock. *Science.* (2013) 342:727–30. doi: 10.1126/science.1243884

52. Yang G, Chen L, Grant GR, Paschos G, Song WL, Musiek ES, et al. Timing of expression of the core clock gene bmal1 influences its effects on aging and survival. *Sci Transl Med.* (2016) 8:324ra16. doi: 10.1126/scitranslmed.aad3305

53. Zhang Y, Fang B, Emmett MJ, Damle M, Sun Z, Feng D, et al. Discrete functions of nuclear receptor Rev-Erb&#X3b1; couple metabolism to the clock. *Science.* (2015) 348:1488–92. doi: 10.1126/science.aab3021

54. Lanier LL. Nk cell receptors. *Annu Rev Immunol.* (1998) 16:359–93. doi: 10.1146/annurev.immunol.16.1.359

55. Posch PE, Hurley CK. Chapter 39 - Histocompatibility: Hla and other systems. In: Porwit A, McCullough J, Erber WN, editors. *Blood and Bone Marrow Pathology (Second Edition).* Edinburgh: Churchill Livingstone (2011). p. 641–76. doi: 10.1016/B978-0-7020-3147-2.00039-0

56. Kollnberger S, Chan A, Sun M-Y, Ye Chen L, Wright C, di Gleria K, et al. Interaction of Hla-B27 homodimers with Kir3dl1 and Kir3dl2, unlike Hla-B27 heterotrimers, is independent of the sequence of bound peptide. *Eur J Immunol.* (2007) 37:1313–22. doi: 10.1002/eji.200635997

57. Ballek O, Valečka J, Dobešová M, Broučková A, Manning J, Rehulka P, et al. Tcr triggering induces the formation of lck–rack1–actinin-1 multiprotein network affecting lck redistribution. *Front Immunol.* (2016) 7:449. doi: 10.3389/fimmu.2016.00449

58. Jiang Y-p, Peng Y-q, Wang L, Qin J, Zhang Y, Zhao Y-z, et al. Rna-sequencing identifies differentially expressed genes in T helper 17 cells in peritoneal fluid of patients with endometriosis. *J Reproduct Immunol.* (2022) 149:103453. doi: 10.1016/j.jri.2021.103453

59. Qiu G, Liu J, Cheng Q, Wang Q, Jing Z, Pei Y, et al. Impaired autophagy and defective T cell homeostasis in mice with T cell-specific deletion of receptor for activated C kinase 1. *Front Immunol.* (2017) 8:575. doi: 10.3389/fimmu.2017.00575

60. Osborne DG, Piotrowski JT, Dick CJ, Zhang J-S, Billadeau DD. Snx17 affects T Cell activation by regulating Tcr and integrin recycling. *J Immunol.* (2015) 194:1402734. doi: 10.4049/jimmunol.1402734

61. Lopez BI, Santiago KG, Lee D, Ha S, Seo K. Rna sequencing (Rna-Seq) based transcriptome analysis in immune response of holstein cattle to killed vaccine against bovine viral diarrhea virus type I. *Animals.* (2020) 10:344. doi: 10.3390/ani10020344

62. McGill JL, Rusk RA, Guerra-Maupome M, Briggs RE, Sacco RE. Bovine gamma delta T cells contribute to exacerbated Il-17 production in response to co-infection with bovine Rsv and mannheimia haemolytica. *PLoS ONE.* (2016) 11:e0151083. doi: 10.1371/journal.pone.0151083

63. Ma WT, Yao XT, Peng Q, Chen DK. The protective and pathogenic roles of Il-17 in viral infections: friend or foe? *Open Biol.* (2019) 9:190109. doi: 10.1098/rsob.190109

64. Merle NS, Church SE, Fremeaux-Bacchi V, Roumenina LT. Complement system part i – molecular mechanisms of activation and regulation. *Front Immunol.* (2015) 6:262. doi: 10.3389/fimmu.2015.00262

65. Ling M, Murali M. Analysis of the complement system in the clinical immunology laboratory. *Clin Lab Med.* (2019) 39:579–90. doi: 10.1016/j.cll.2019.07.006

66. Kurtovic L, Beeson JG. Complement factors in Covid-19 therapeutics and vaccines. *Trends Immunol.* (2021) 42:94–103. doi: 10.1016/j.it.2020.12.002

67. Kim YJ, Kim KH, Ko EJ, Kim MC, Lee YN, Jung YJ, et al. Complement C3 plays a key role in inducing humoral and cellular immune

responses to influenza virus strain-specific hemagglutinin-based or cross-protective M2 extracellular domain-based vaccination. *J Virol.* (2018) 92:e00969–18. doi: 10.1128/JVI.00969-18

68. Carroll MC. The complement system in regulation of adaptive immunity. *Nat Immunol.* (2004) 5:981–6. doi: 10.1038/ni1113

69. Molina H, Kinoshita T, Webster CB, Holers VM. Analysis of C3b/C3d binding sites and factor I cofactor regions within mouse complement receptors 1 and 2. *J Immunol.* (1994) 153:789–95.

70. Kopf M, Abel B, Gallimore A, Carroll M, Bachmann MF. Complement component C3 promotes T-cell priming and lung migration to control acute influenza virus infection. *Nat Med.* (2002) 8:373–8. doi: 10.1038/nm0402-373

71. Salehen Na, Stover C. The role of complement in the success of vaccination with conjugated vs. unconjugated polysaccharide antigen. *Vaccine.* (2008) 26:451–9. doi: 10.1016/j.vaccine.2007.11.049

72. Geering B, Stoeckle C, RoŽman S, Oberson K, Benarafa C, Simon H-U. Dapk2 positively regulates motility of neutrophils and eosinophils in response to intermediary chemoattractants. *J Leukoc Biol.* (2014) 95:293–303. doi: 10.1189/jlb.0813462

73. Ni X, Wang Y, Wang P, Chu C, Xu H, Hu J, et al. Death associated protein kinase 2 suppresses T-B interactions and Gc formation. *Mol Immunol.* (2020) 128:249–57. doi: 10.1016/j.molimm.2020.10.018

74. Kim H, Dickey L, Stone C, Jafek JL, Lane TE, Tantin D, et al. Cell-selective deletion of Oct1 protects animals from autoimmune neuroinflammation while maintaining neurotropic pathogen response. *J Neuroinflammation.* (2019) 16:133. doi: 10.1186/s12974-019-1523-3

75. Shakya A, Goren A, Shalek A, German CN, Snook J, Kuchroo VK, et al. Oct1 and Oca-B are selectively required for Cd4 memory T cell function. *J Exp Med.* (2015) 212:2115–31. doi: 10.1084/jem.20150363

76. Hwang SS, Kim LK, Lee GR, Flavell RA. Role of Oct-1 and partner proteins in T cell differentiation. *Biochim Biophys Acta.* (2016) 1859:825–31. doi: 10.1016/j.bbagrm.2016.04.006

77. Prado DS, Cattley RT, Shipman CW, Happe C, Lee M, Boggess WC, et al. Synergistic and additive interactions between receptor signaling networks drive the regulatory T cell versus T helper 17 cell fate choice. *J Biol Chem.* (2021) 297:101330. doi: 10.1016/j.jbc.2021.101330

78. Gao Y, Zhang X, Wang Z, Qiu Y, Liu Y, Shou S, et al. The contribution of neuropilin-1 in the stability of Cd4+Cd25+ regulatory T cells through the Tgf-?1/smads signaling pathway in the presence of lipopolysaccharides. *Immun Inflamm Dis.* (2022) 10:143–54. doi: 10.1002/iid3.551

79. Ran Y, Zhang J, Liu LL, Pan ZY, Nie Y, Zhang HY, et al. Autoubiquitination of trim26 links Tbk1 to nemo in Rlr-mediated innate antiviral immune response. *J Mol Cell Biol.* (2016) 8:31–43. doi: 10.1093/jmcb/mjv068

80. Huang H, Sharma M, Zhang Y, Li C, Liu K, Wei J, et al. Expression profile of porcine Trim26 and its inhibitory effect on interferon-β production and antiviral response. *Genes.* (2020) 11:1226. doi: 10.3390/genes11101226

81. Zhao J, Cai B, Shao Z, Zhang L, Zheng Y, Ma C, et al. Trim26 positively regulates the inflammatory immune response through k11-linked ubiquitination of tab1. *Cell Death Differ.* (2021) 28:3077–91. doi: 10.1038/s41418-021-00803-1

82. Dargaud Y, Scoazec JY, Wielders SJH, Trzeciak C, Hackeng TM, Négrier C, et al. Characterization of an autosomal dominant bleeding disorder caused by a thrombomodulin mutation. *Blood.* (2015) 125:1497–501. doi: 10.1182/blood-2014-10-604553

83. Adams TE, Huntington JA. Thrombin-cofactor interactions. *Arterioscler Thromb Vasc Biol.* (2006) 26:1738–45. doi: 10.1161/01.ATV.0000228844.65168.d1

84. Millar DG, Garza KM, Odermatt B, Elford AR, Ono N, Li Z, et al. Hsp70 promotes antigen-presenting cell function and converts T-Cell tolerance to autoimmunity in vivo. *Nat Med.* (2003) 9:1469–76. doi: 10.1038/nm962

85. Castellino F, Boucher PE, Eichelberg K, Mayhew M, Rothman JE, Houghton AN, et al. Receptor-mediated uptake of antigen/heat shock protein complexes results in major histocompatibility complex class I antigen presentation via two distinct processing pathways. *J Exp Med.* (2000) 191:1957–64. doi: 10.1084/jem.191.11.1957

86. Ishii T, Udono H, Yamano T, Ohta H, Uenaka A, Ono T, et al. Isolation of Mhc class I-restricted tumor antigen peptide and its precursors associated with heat shock proteins Hsp70, Hsp90, and Gp96. *J Immunol.* (1999) 162:1303–9.

87. Asea A, Kraeft SK, Kurt-Jones EA, Stevenson MA, Chen LB, Finberg RW, et al. Hsp70 stimulates cytokine production through a Cd14-dependant pathway, demonstrating its dual role as a chaperone and cytokine. *Nat Med.* (2000) 6:435–42. doi: 10.1038/74697

88. Ambade A, Catalano D, Lim A, Mandrekar P. Inhibition of heat shock protein (molecular weight 90 Kda) attenuates proinflammatory cytokines and prevents lipopolysaccharide-induced liver injury in mice. *Hepatology.* (2012) 55:1585–95. doi: 10.1002/hep.24802

89. Ambade A, Catalano D, Lim A, Kopoyan A, Shaffer SA, Mandrekar P. Inhibition of heat shock protein 90 alleviates steatosis and macrophage activation in murine alcoholic liver injury. *J Hepatol.* (2014) 61:903–11. doi: 10.1016/j.jhep.2014.05.024

90. Lin L, Pan S, Zhao J, Liu C, Wang P, Fu L, et al. Hspd1 Interacts with Irf3 to facilitate interferon-beta induction. *PLoS ONE.* (2014) 9:e114874. doi: 10.1371/journal.pone.0114874

91. Mo Z, Zhang S, Zhang S. A novel signature based on mtorc1 pathway in hepatocellular carcinoma. *J Oncol.* (2020) 2020:8291036. doi: 10.21203/rs.3.rs-18272/v1

92. Cwiklinska H, Cichalewska-Studzinska M, Selmaj KW, Mycko MP. The heat shock protein Hsp70 promotes Th17 genes' expression via specific regulation of microrna. *Int J Mol Sci.* (2020) 21:2823. doi: 10.3390/ijms21082823

93. Yang Y, Xia J, Yang Z, Wu G, Yang J. The abnormal level of Hsp70 is related to Treg/Th17 imbalance in pcos patients. *J Ovarian Res.* (2021) 14:155. doi: 10.1186/s13048-021-00867-0

94. Kim BK, Park M, Kim JY, Lee KH, Woo SY. Heat shock protein 90 is involved in Il-17-mediated skin inflammation following thermal stimulation. *Int J Mol Med.* (2016) 38:650–8. doi: 10.3892/ijmm.2016.2627

95. Baldwin CL, Damani-Yokota P, Yirsaw A, Loonie K, Teixeira AF, Gillespie A. Special features of *Γ*δ T cells in ruminants. *Mol Immunol.* (2021) 134:161–9. doi: 10.1016/j.molimm.2021.02.028

96. Houchins JP, Yabe T, McSherry C, Bach FH. DNA sequence analysis of Nkg2, a family of related cdna clones encoding type ii integral membrane proteins on human natural killer cells. *J Exp Med.* (1991) 173:1017–20. doi: 10.1084/jem.173.4.1017

97. Guzman E, Birch JR, Ellis SA. Cattle mic is a ligand for the activating Nk Cell receptor Nkg2d. *Vet Immunol Immunopathol.* (2010) 136:227–34. doi: 10.1016/j.vetimm.2010.03.012

98. Blunt MD, Vallejo Pulido A, Fisher JG, Graham LV, Doyle ADP, Fulton R, et al. Kir2ds2 expression identifies Nk cells with enhanced anticancer activity. *J Immunol.* (2022) 209:379–90. doi: 10.4049/jimmunol.2101139

99. Closa L, Xicoy B, Zamora L, Estrada N, Colomer D, Herrero MJ, et al. Natural killer cell receptors and ligand variants modulate response to tyrosine kinase inhibitors in patients with chronic myeloid leukemia. *Hla.* (2022) 99:93–104. doi: 10.1111/tan.14515

100. Moftah NH, El-Barbary RA, Rashed L, Said M. Ulbp3: a marker for alopecia areata incognita. *Arch Dermatol Res.* (2016) 308:415–21. doi: 10.1007/s00403-016-1652-9

101. Sutherland CL, Rabinovich B, Chalupny NJ, Brawand P, Miller R, Cosman D. Ulbps, human ligands of the Nkg2d receptor, stimulate tumor immunity with enhancement by Il-15. *Blood.* (2006) 108:1313–9. doi: 10.1182/blood-2005-11-011320

102. Mou X, Zhou Y, Jiang P, Zhou T, Jiang Q, Xu C, et al. The regulatory effect of Ul-16 binding protein-3 expression on the cytotoxicity of Nk cells in cancer patients. *Sci Rep.* (2014) 4:6138. doi: 10.1038/srep06138

103. Bassi PB, Araujo FF, Garcia GC, Costa ESMF, Bittar ER, Bertonha CM, et al. Haematological and immunophenotypic evaluation of peripheral blood cells of cattle naturally infected with bovine papillomavirus. *Vet J.* (2019) 244:112–5. doi: 10.1016/j.tvjl.2018.12.004

104. Hamilton CA, Mahan S, Entrican G, Hope JC. Interactions between natural killer cells and dendritic cells favour T helper1-type responses to bcg in calves. *Vet Res.* (2016) 47:85. doi: 10.1186/s13567-016-0367-4

105. De Beuckelaer A, Pollard C, Van Lint S, Roose K, Van Hoecke L, Naessens T, et al. Type I interferons interfere with the capacity of Mrna lipoplex vaccines to elicit cytolytic T cell responses. *Mol Ther.* (2016) 24:2012–20. doi: 10.1038/mt.2016.161

106. Zhong C, Liu F, Hajnik RJ, Yao L, Chen K, Wang M, et al. Type I interferon promotes humoral immunity in viral vector vaccination. *J Virol.* (2021) 95:e0092521. doi: 10.1128/JVI.00925-21

107. Ye L, Ohnemus A, Ong LC, Gad HH, Hartmann R, Lycke N, et al. Type I and type iii interferons differ in their adjuvant activities for influenza vaccines. *J Virol.* (2019) 93:e01262–19. doi: 10.1128/JVI.01262-19

108. Crow MK, Ronnblom L. Type I interferons in host defence and inflammatory diseases. *Lupus Sci Med.* (2019) 6:e000336. doi: 10.1136/lupus-2019-000336

109. Tizioto PC, Kim J, Seabury CM, Schnabel RD, Gershwin LJ, Van Eenennaam AL, et al. Immunological response to single pathogen challenge with agents of the bovine respiratory disease complex: an rna-sequence analysis of the bronchial lymph node transcriptome. *PLoS ONE.* (2015) 10:e0131459. doi: 10.1371/journal.pone.0131459

110. Behura SK, Tizioto PC, Kim J, Grupioni NV, Seabury CM, Schnabel RD, et al. Tissue tropism in host transcriptional response to members of the bovine respiratory disease complex. *Sci Rep.* (2017) 7:17938. doi: 10.1038/s41598-017-18205-0

111. Johnston D, Earley B, McCabe MS, Kim J, Taylor JF, Lemon K, et al. Messenger Rna biomarkers of bovine respiratory syncytial virus infection in the whole blood of dairy calves. *Sci Rep.* (2021) 11:9392. doi: 10.1038/s41598-021-88878-1

112. Scott MA, Woolums AR, Swiderski CE, Perkins AD, Nanduri B. Genes and regulatory mechanisms associated with experimentally-induced bovine respiratory disease identified using supervised machine learning methodology. *Sci Rep.* (2021) 11:22916. doi: 10.1038/s41598-021-02343-7

TYPE Review
PUBLISHED 11 November 2022
DOI 10.3389/fvets.2022.1008728

OPEN ACCESS

EDITED BY
Joanna Szyda,
Wroclaw University of Environmental and Life Sciences, Poland

REVIEWED BY
Saurabh Gupta,
GLA University, India
Mahmoud Madkour,
National Research Centre, Egypt

*CORRESPONDENCE
Jun-Mo Kim
junmokim@cau.ac.kr

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

RECEIVED 01 August 2022
ACCEPTED 31 October 2022
PUBLISHED 11 November 2022

CITATION
Pathak RK and Kim J-M (2022) Vetinformatics from functional genomics to drug discovery: Insights into decoding complex molecular mechanisms of livestock systems in veterinary science. *Front. Vet. Sci.* 9:1008728. doi: 10.3389/fvets.2022.1008728

Vetinformatics from functional genomics to drug discovery: Insights into decoding complex molecular mechanisms of livestock systems in veterinary science

Rajesh Kumar Pathak and Jun-Mo Kim*

Department of Animal Science and Technology, Chung-Ang University, Anseong-si, South Korea

Having played important roles in human growth and development, livestock animals are regarded as integral parts of society. However, industrialization has depleted natural resources and exacerbated climate change worldwide, spurring the emergence of various diseases that reduce livestock productivity. Meanwhile, a growing human population demands sufficient food to meet their needs, necessitating innovations in veterinary sciences that increase productivity both quantitatively and qualitatively. We have been able to address various challenges facing veterinary and farm systems with new scientific and technological advances, which might open new opportunities for research. Recent breakthroughs in multi-omics platforms have produced a wealth of genetic and genomic data for livestock that must be converted into knowledge for breeding, disease prevention and management, productivity, and sustainability. Vetinformatics is regarded as a new bioinformatics research concept or approach that is revolutionizing the field of veterinary science. It employs an interdisciplinary approach to understand the complex molecular mechanisms of animal systems in order to expedite veterinary research, ensuring food and nutritional security. This review article highlights the background, recent advances, challenges, opportunities, and application of vetinformatics for quality veterinary services.

KEYWORDS
vetinformatics, livestock systems, functional genomics, drug discovery, veterinary science

Introduction

Livestock animals are an essential part of our life. Science-led innovation in veterinary research that benefits both people and animals as individuals and populations is crucial to maintaining public health (1, 2). This encompasses research on fundamental animal biology and animal welfare, as well as disease prevention, diagnosis, and therapy. Such innovation offers several opportunities for improving animal and human health (3, 4). Currently, veterinarians face many challenges exacerbated by climate change, including the emergence of new diseases, as well as those of a rapidly growing human population that requires adequate food and nutrition. Therefore, integration of interdisciplinary approaches with veterinary science is urgently needed to decode the complex molecular mechanisms of livestock systems (5–7).

The functioning of livestock systems is an area of active, ongoing research. Advancements in mathematical science, statistical methods, computer science, and information technology help biologists learn about biological systems quantitatively and qualitatively (8). Computers are essential components of these scientific advancements, as they play a crucial role in research and development sectors and become a major tool for researchers. In the era of "omics," we can easily handle big data using computers, but the term "bioinformatics" was not introduced until the beginning of the 1970s by Hogeweg and Ben Hesper, when DNA could not yet be sequenced (9, 10). DNA's role as genetic material was also a matter of debate before 1952. Avery et al. (11) demonstrated that a non-virulent bacterial strain could acquire virulence by absorbing purified DNA from a virulent strain (8). However, the scientific community did not immediately accept their findings. Many scientists instead believed that proteins, rather than DNA, were carriers of genetic information (8, 12). Hershey and Chase established the role of DNA as a genetic information–encoding molecule in 1952 when they demonstrated that bacteriophage-infected bacterial cells ingest and transfer DNA rather than protein (13). At this time, DNA's primary role was understood, but little was known about how the DNA molecule was arranged. It was only known that its monomers (i.e., nucleotides) were present proportionately (14). The DNA double-helix structure was finally discovered by Watson and Crick (15). Despite this achievement, it would still be another 13 years before the genetic code was cracked, and another 25 years before the first DNA sequencing techniques were made accessible (16–18). Accordingly, DNA analysis using computational tools lagged ∼2 decades behind the study of proteins, whose chemical makeup was already better understood than that of DNA (8).

Due to significant improvements in the crystallographic determination of protein structures (19), protein analysis was bioinformatics' starting point in Gauthier et al. (8). Insulin's sequence, or the arrangement of its amino acids, was the first protein sequence to be published (20). Additionally, numerous improvements in determining the structure and sequence of proteins were also reported (10, 21). The first bioinformatician was an American physical chemist named Margaret Dayhoff (1925–1983) who made significant contributions and used computational approaches in the study of biochemistry and protein sciences. She is referred to as the mother and father of bioinformatics (19, 20, 22).

Needleman and Wunsch created the first dynamic programming method for pairwise protein sequence alignment in 1978 (23). Since the early 1980s, multiple sequence alignment (MSA) algorithms have been emerging, facilitated by CLUSTAL software, which was introduced to MSA in 1988 (24, 25). Further, the concept of a mathematical framework for amino acid substitution was introduced by Dayhoff with the development of a point accepted mutation matrix (26). In the 1970s, DNA became more actively researched than proteins. Additionally, parallel developments in biology and computer science took place in the 1980 and 1990. Since the establishment of the National Center for Biotechnology Information (NCBI) in 1988 and the start of the Human Genome Project in 1990, bioinformatics has received significant attention and become an integral part of the analysis of the human genome (27–29). Further, bioinformatics emerged as a separate interdisciplinary subject for research and development in different areas of science and technology (10). Its approaches are extensively utilized in biomedical and pharmaceutical research. In recent years, the veterinary science community has sought to use these approaches in their research. Therefore, the concept of vetinformatics has been introduced as a branch of bioinformatics that focuses on livestock animals for quality veterinary services (5, 30).

In veterinary science, the livestock production system is a complex process that has three interconnected basic components: animal biology, the environment, and management techniques (31). Therefore, vetinformatics approaches are required to bridge the gaps between genotype and phenotype in order to improve livestock productivity and sustainability (5, 30). Large animal datasets have been produced as a result of improvements in various omics platforms and next-generation sequencing (NGS) technologies. Several bioinformatics databases and tools are available for their management and analysis, but these databases hold information about diverse groups of organisms (7, 10), and veterinarians require species-specific databases. Additionally, they require animal-based vetinformatics tools for data analysis and integration, as well as computational and mathematical models for analyzing animal behavior (32, 33). Accordingly, vetinformatics is required as a separate interdisciplinary subject to handle livestock data. By analyzing these large data sets, it is possible to accelerate research and development by extracting crucial information that enables researchers to understand livestock systems at molecular levels (Figure 1).

Scientific disciplines linked with vetinformatics and their support systems

Vetinformatics is associated with the disciplines of veterinary sciences, basic sciences, and engineering; these

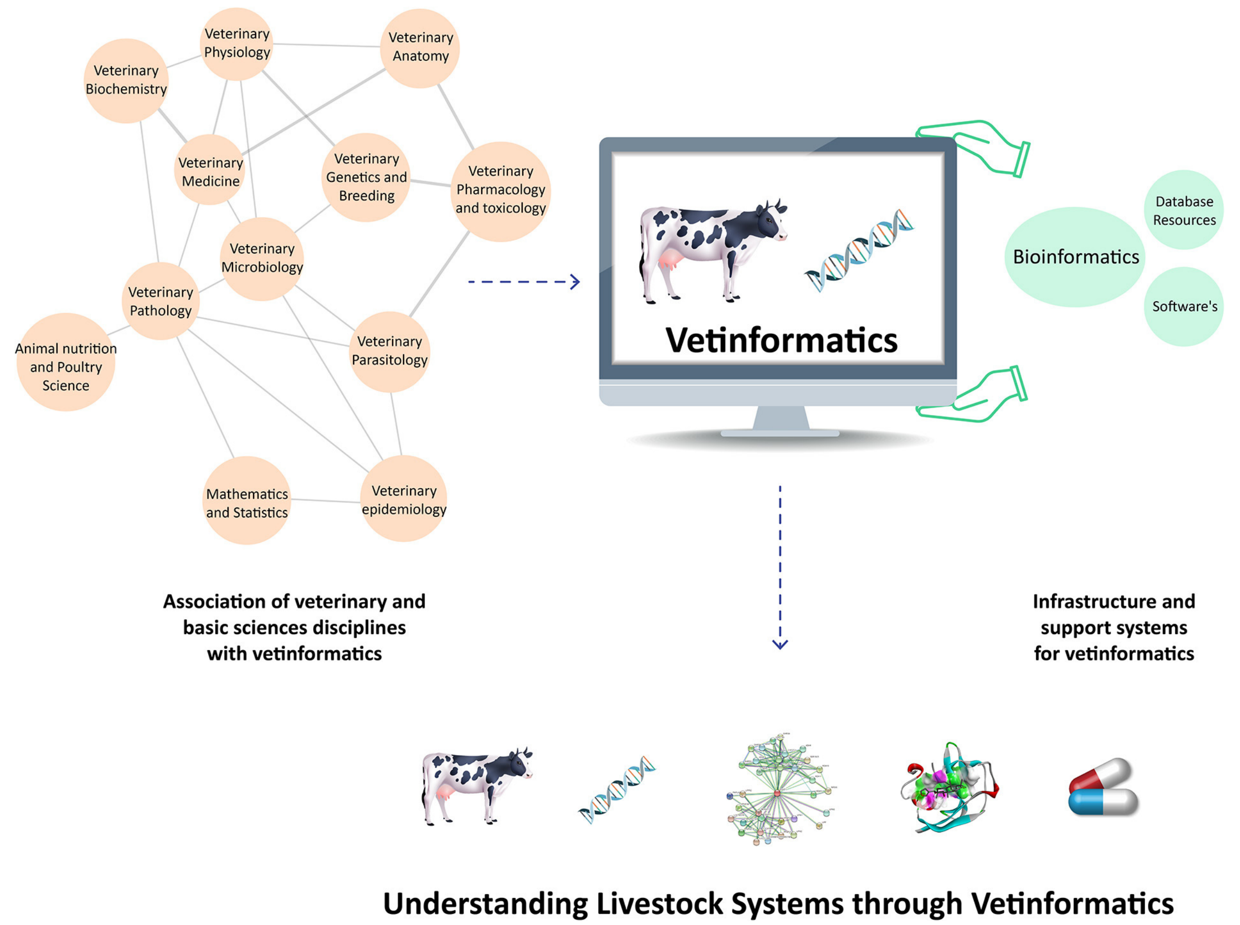

Veterinary Physiology
Veterinary Anatomy
Veterinary Biochemistry
Veterinary Medicine
Veterinary Genetics and Breeding
Veterinary Pharmacology and toxicology
Veterinary Microbiology
Veterinary Pathology
Veterinary Parasitology
Animal nutrition and Poultry Science
Mathematics and Statistics
Veterinary epidemiology
Vetinformatics
Bioinformatics
Database Resources
Software's
Association of veterinary and basic sciences disciplines with vetinformatics
Infrastructure and support systems for vetinformatics
Understanding Livestock Systems through Vetinformatics

FIGURE 1
Integration of veterinary sciences, basic science disciplines, and support systems to create vetinformatics, enabling a better understanding of livestock systems in veterinary science.

disciplines provide infrastructure and an interdisciplinary nature to vetinformatics (5, 30, 33, 34). Several traditional and advanced subjects are associated with vetinformatics, such as veterinary physiology, biochemistry, anatomy, pharmacology and toxicology, parasitology, microbiology, pathology, epidemiology, genetics and breeding, and medicine, as well as animal nutrition and poultry science. Mathematical and statistical sciences also contribute to vetinformatics as basic-science disciplines. Computer science, information technology (IT), and computational resources serve as the foundation and support system for vetinformatics (5, 34). Accordingly, vetinformatics is created through the integration of veterinary sciences, basic sciences, and supporting disciplines. Vetinformatics uses computer science and IT to quickly provide solutions to complex challenges related to livestock systems (https://www.frontiersin.org/research-topics/33198/vetinformatics-an-insight-for-decoding-livestock-systems-through-in-silico-biology).

Needs and aims of vetinformatics

In order to better understand livestock systems, vetinformatics is expanding and has contributed to the growth of research initiatives involving high-throughput DNA sequencing data analysis and other omics fields (https://www.veterinaryirelandjournal.com/ucd-research/165-how-omics-are-contributing-to-sustainable-animal-production; accessed on 14/7/2022). Vetinformatics' aim is to decode the enormous quantity of multi-omics data produced by high-throughput technologies, structural and functional characterizations of key genes and proteins, and visualizations of key components linked with livestock productivity and sustainability (7).

In general, the goals of vetinformatics include building databases that document information on medicinal plants, particularly for the development of herbal veterinary medicines *via* screening of phytochemicals against molecular drug targets using molecular docking (5, 6, 35). However, vetinformatics' aims also include collecting animal genetic resources in databases, developing these databases for managing omics data sets that are species- or organism-specific, and enhancing the content of existing veterinary databases so that they can be used more effectively (36–38). In addition, developing platform-independent graphical user interface–based software for integration and analysis of multi-omics data (10, 32–34), and improving the accessibility of public software for veterinary biotechnologists, scientists, and veterinarians would further vetinformatics' missions. Finally, vetinformatics also strives to educate undergraduate students, graduate students, and faculty of veterinary and animal sciences about the use of vetinformatics for the analysis of multi-omics data [(32, 34) https://www.frontiersin.org/research-topics/33198/vetinformatics-an-insight-for-decoding-livestock-systems-through-in-silico-biology].

Recent advances in vetinformatics

The integration of omics science and technology with veterinary science opens exciting opportunities to decode livestock systems (7, 36, 39). Many animal genomes have been sequenced, and others are currently under sequencing and analysis. Although multi-omics data are generated regularly, more research is still needed to increase the efficiency and standardize interpretation, analysis, and integration of these data (7, 40). Additionally, the availability of big data in veterinary science has helped in the design of innovative algorithms and improved knowledge of cellular and phenotypic mechanisms [(7) https://ivcjournal.com/ai-the-newest-tool-in-veterinary-science/; accessed July 14, 2022]. Some animal-specific databases have already been developed to provide updated information to the veterinary science community (4, 37). Further, the use of machine and deep learning approaches in livestock research is reshaping the field in unanticipated ways. Exciting, cutting-edge models that connect genotype and phenotype allow the field of vetinformatics to grow quickly in the digital era and improve livestock productivity (41, 42).

Challenges in vetinformatics

Vetinformatics is connected to veterinary sciences, basic sciences, and technology. Due to its interdisciplinary nature, vetinformatics may include people with a background in veterinary science or biology with little interest in computer programming, or people with a background in computer science who are unfamiliar with certain biological concepts. Due to the importance and application of vetinformatics in livestock research, many post-graduate programs in veterinary and animal science require exposure to the subject of vetinformatics. These programs may develop student interest in this emerging and interdisciplinary field, filling an urgent need for more researchers in vetinformatics. The major challenges faced by vetinformatics include managing big data in veterinary sciences; developing species-specific databases and tools for livestock research; improving the accuracy of available tools; developing novel algorithms and tools; analyzing and integrating multi-omics data; and identifying molecules for the development of drugs for treating livestock diseases (43, 44).

Applications of vetinformatics in health, productivity, and sustainability of livestock

The scientific community produces complex data daily by using advanced molecular biology and biotechnology-based techniques (45, 46). These techniques require statistical approaches to quickly and accurately interpret these large-scale data (7, 40). Computational studies are the only method for analysis and interpretation of genome sequencing, assembly and alignment, differentially expressed genes, biological networks, protein modeling as well as molecular docking. The integration of such data is made possible by statistical and mathematical modeling approaches (5, 6, 10, 47, 48). Vetinformatics has tremendous potential to address challenging issues in veterinary science and related fields. Today, it is a vital tool for scientists and is crucial to the study of livestock. The following sections highlight the applications of vetinformatics.

Assembly and annotation of newly sequenced genomes

Sequencing an animal's genome is necessary to understand the intricacy among them (49, 50). Aligning and combining fragments of genome sequences obtained from sequencing platforms is referred to as assembly. Depending on whether a reference genome is available or not, assembly can be divided into two categories: *de novo* assembly and reference-based assembly (50, 51). Genome assembly is essential for determining how gene structure and function will affect an organism's behavior. SPAdes is a highly cited genome assembler originally designed for small genomes. It was tested on microorganisms including bacteria, fungi, and other small genomes (52). Besides, it includes various assembly pipelines such as metaSPAdes, plasmidSPAdes, rnaSPAdes, truSPAdes, and dipSPAdes. These pipelines are useful for metagenomic data sets, assembly of plasmids from WGS data, *de novo* assembly of RNA-Seq data, barcode assembly, and highly polymorphic diploid genomes (https://cab.spbu.ru/files/release3.12.0/manual.html). In the field of genomics, annotating genomes through MAKER is convenient and easy. Genomes of eukaryotes and prokaryotes can be annotated independently and genome databases can be created using it. It is designed to identify repeats, align ESTs and proteins with the genomes, and produce *ab-initio* gene predictions (53).

Using high-throughput sequencing platforms, we now have sequenced genomes for many major animal species (54, 55). Zimin et al. (56) used a combination of whole-genome shotgun sequencing and hierarchical sequencing techniques to sequence the genome of the domestic cow (*Bos taurus*). They assembled the 35 million sequence reads to produce an improved assembly of 2.86 billion base pairs. Numerous computational tools can be used to further evaluate sequenced genomes. Several pipelines, resources, and software are available for computational assembly and study of the genome. Recent functional annotation of three domestic animal genomes (cattle, chicken, and pig) provides a useful resource for livestock research (57). The study of the genome is pertinent to many areas of research, including ancestry determination, genomic selection, and vaccine and drug design (58–60).

Transcriptome and RNA-Seq data analysis for studying gene expression

RNA-Seq has emerged as an effective approach for transcriptome analyses that will eventually make microarrays outdated for analysis of gene expression data (61). The field of research on gene expression has undergone a recent revolution. This technology has made possible the measure of simultaneous gene expression, enabling the discovery of candidate genes with potential biological significance (62). RNA-Seq analysis of porcine ovaries revealed 4,414 deferentially expressed genes and helped to discover their roles in the late metestrus and diestrus phases of the estrous cycle (48). The findings from a separate transcriptome analysis strongly suggested that IGF1, PGR, ITPR1, and CHRM3 regulate oocyte maturation and smooth muscle contraction in pigs, and provided direction for future research involving effective animal breeding programs (63).

Non-coding RNAs such as small interfering RNAs (siRNAs) and microRNAs (miRNA) play vital roles in gene regulation (64). Recent investigations have demonstrated that they are effective in treating a variety of diseases, and working as a biomarker for effective therapies (64–66). The role of miRNAs has been examined in several studies with respect to livestock diseases (66). Several candidate genes and miRNAs have been identified that could be helpful in treating mastitis disease in cows (67, 68). Vetinformatics-based approaches are useful in detection of siRNAs in host and their targets in pathogen genomes, leading to the development of novel treatments against livestock diseases (64, 69).

Metagenomic analysis for dissecting microbial communities and their role in livestock

Metagenomics allows for direct access to genetic information of whole communities by utilizing a variety of genomic technologies and computational approaches (39). It presents a considerably more comprehensive description than phylogenetic surveys because it enables access to the functional gene composition of microbial communities. Metagenomics provides information on potentially novel enzymes or biocatalysts, relationships between phylogeny and function for uncultured organisms, and evolutionary profiles of community structure and function (39). Kumar et al. (70) analyzed metagenomic data of bacterial communities in pig slurries, enhancing knowledge of how microbial abundance in swine slurries varies over time. Another microbiome analysis characterized changes in microbial community composition that resulted from feeding dairy cows one of two common diets: pasture and total mixed ration. Studies such as this one will contribute to the management of cattle feed and the study of rumen microbial ecology (71). On larger scales, metagenomics-based analyses will help to improve animal health, leading to enhanced livestock productivity and sustainability.

Sequence alignment and analysis for identification of biologically significant regions

With the availability of the BLAST tool beginning in 1990, sequence analysis has emerged as a key area of research (72). The field of sequence analysis is fairly broad, but in this section, we will focus on the analysis of nucleotide or protein sequences. A variety of sequence-alignment tools such as BLAST, FASTA, and Muscle are available to identify or compare two (pair-wise alignment) or more (multiple sequence alignment; MSA) sequences (10). Ajayi et al. (73) identified 67 genes in the bovine genome belonging to heat shock protein families using sequence alignment and analysis. Using *in silico* analysis, the study investigated transcription start sites and promoter regions of olfactory receptors in cattle, identifying five candidate motifs (MOR1, MOR2, MOR3, MOR4, and MOR5) important in gene regulation (74). It is also used to annotate recently discovered sequences, identify conserved and regulatory regions, and predict sequence physicochemical properties (10).

Molecular phylogeny for analyzing relationship among organisms

Another crucial area of research in vetinformatics is molecular phylogenetic analysis (75). Widely employed in evolutionary biology, molecular phylogenetic analysis can identify similarities between various animal sequences in order to infer their evolutionary relationship (76). Additionally, it facilitates the identification of critical elements in individual sequences and their association with other sequences, thereby playing an important role in drug and vaccine design (10, 76). In the field of molecular phylogeny, the Molecular Evolutionary Genetic Analysis (MEGA) (77) and PHYLIP (78) are well-known software. Besides, several other web-based tools are also available such as Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/), and T-Coffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/) to perform multiple sequence alignment and building a phylogenetic tree using different methods. There has been increasing interest in reconstructing phylogenetic trees in biological science, and questions are being raised regarding the degree of confidence one should place in any given phylogenetic tree. The concept of bootstrapping and jackknifing was introduced to construct error-free phylogenetic tree (79). Phylogenetic analysis software like MEGA facilitates researchers to set a bootstrap value during phylogenetic tree reconstruction to confirm their accuracy (https://www.megasoftware.net/web_help_11/Bootstrap_Test_of_Phylogeny.htm). The Interactive Tree Of Life, i.e., iTOL (https://itol.embl.de/) and Tree View are highly cited tools facilitating phylogenetic tree visualization (80, 81). The bovine hepacivirus (BovHepV) of five positive samples that formed a separate branch from other BovHepV in a phylogenetic analysis conducted by Deng et al. based on the partial NS3 coding sequence (82). The findings suggested that these new BovHepV represent novel and emerging strains. Another study that conducted a molecular characterization and phylogenetic analysis of the lumpy skin disease virus (LSDV) that is circulating in northern Thailand revealed a relationship with other LSDVs. The LSDV that was isolated from northern Thailand shared genetic traits with the LSDVs that are currently circulating in China, Hong Kong, and Vietnam. This finding will be instrumental in developing disease control strategies against LSDVs (83).

Genome wide association study for identification of important genomic regions

Genome-wide association study, commonly known as GWAS, is a powerful approach used to identify genetic variants linked to increased likelihood of a certain disease or trait (84, 85). The approach requires examining a large number of individual genomes in search of genetic variants that are more prevalent in individuals with a particular disease or trait. Once such genetic variants have been discovered, they are often utilized to look for neighboring variants that directly contribute to the disease or trait (84, 85). An analysis of GWAS can be conducted using single-locus, and multi-locus models (86). The General linear model (GLM), Mixed linear model (MLM), Logistic mixed model (LMM), and Compressed mixed linear model (CMLM) are the single locus models (87–89), and multi-locus model includes Multilocus random SNP effect mixed linear models (mrMLM) and Fast multilocus random SNP effect efficient mixed-model association (FASTmrEMMA) (86, 90–92). The computer programs commonly used for GWAS include PLINK (93), GenABEL (94), GenAMap (95), and GEMMA (96). In addition, the genomic databases and genome browsers such as NCBI (https://www.ncbi.nlm.nih.gov/), Animal QTLdb (https://www.animalgenome.org/cgi-bin/QTLdb/index), NAGRP (https://www.animalgenome.org/), Ensembl (https://asia.ensembl.org/index.html), and UCSC (https://genome.ucsc.edu/) are valuable resources (86). Besides, the Genome Analysis Toolkit (GATK) pipeline is an important platform for high-throughput genomics data analysis (97). There are a variety of tools available through GATK, most of which are focused on discovering variants and used for genotyping (https://gatk.broadinstitute.org/hc/en-us). Uemoto et al. identified six significant quantitative trait loci for immune-related traits in pigs affected by mycoplasma pneumonia of swine using GWAS, revealing novel insights into the genomic elements influencing pig production, respiratory illness, and immune-related traits (98). Another GWAS-based study identified candidate genes for milk production traits in Korean Holstein cattle and individual birth weight traits in Korean Yorkshire pigs (99, 100). Therefore, GWAS-based approaches have potential to decode important and complex traits linked with livestock productivity.

Systems biology and integration of multi-omics data

Systems biology is a key subfield of vetinformatics and has made great contributions to the modeling and simulation of biological systems (101–103). The field aids in the integration of multi-omics data, including genomics, proteomics, metabolomics, and transcriptomics, in order to construct models that comprehensively characterize the behavior of biological systems under various conditions (104). In the past, researchers were forced to focus on single genes or proteins, but as omics technology and systems biology have advanced, the paradigm has changed from a reductionist approach to a holistic approach (104). Through network modeling and analysis, systems biology enables prediction of the behavior of whole systems and identification of essential components involved in various biological processes, both of which ultimately contribute to advancements in animal welfare and livestock productivity (101–104).

Network biology and analysis

In network analysis, networks represent relationships among the components of a given system (104, 105). In biological systems, these relationships have attracted

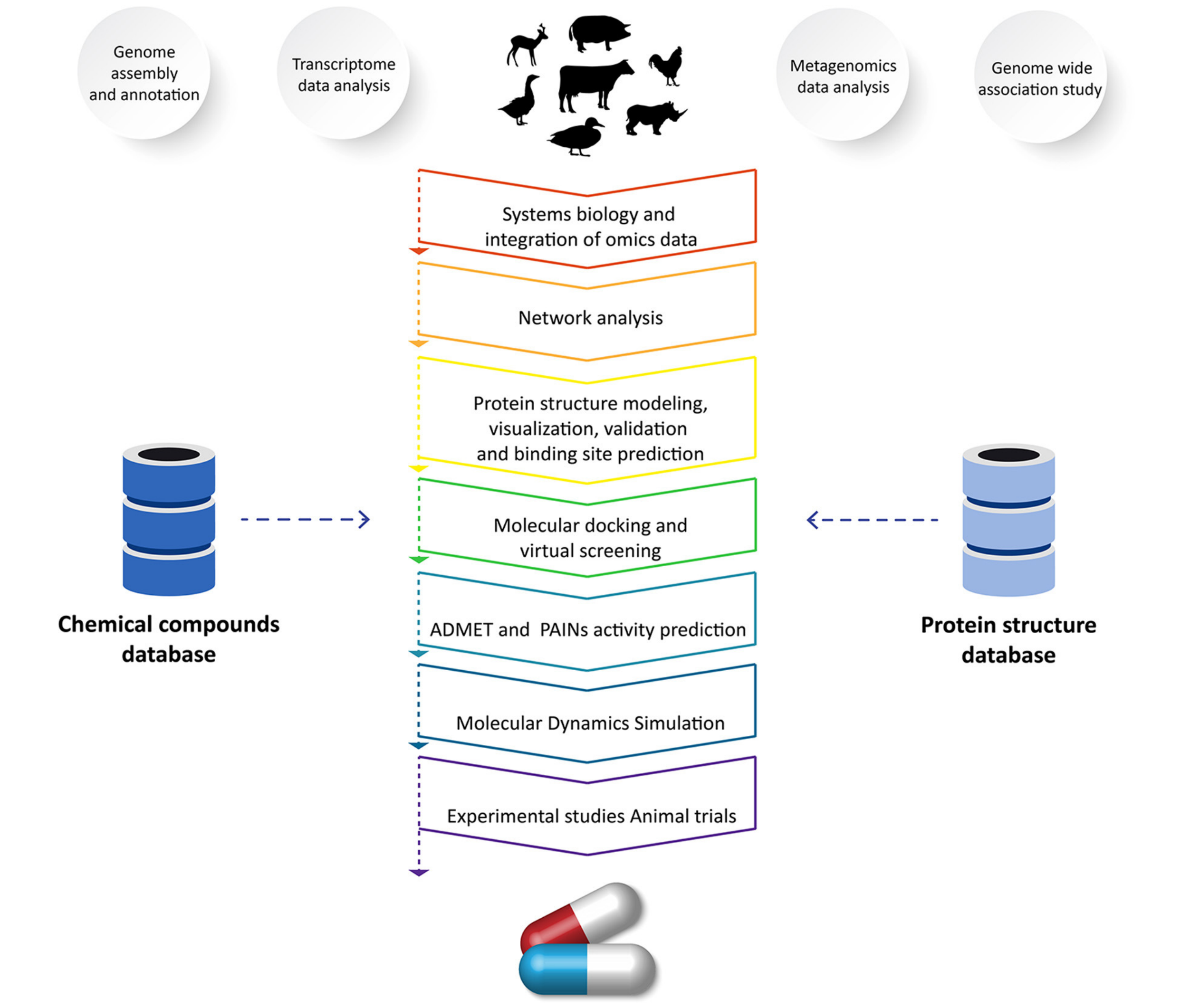

Genome assembly and annotation
Transcriptome data analysis
Metagenomics data analysis
Genome wide association study
Systems biology and integration of omics data
Network analysis
Protein structure modeling, visualization, validation and binding site prediction
Molecular docking and virtual screening
ADMET and PAINs activity prediction
Molecular Dynamics Simulation
Experimental studies Animal trials
Chemical compounds database
Protein structure database

FIGURE 2
Application of vetinformatics to analyze high-throughput sequencing data for discovery of novel drug molecule(s) for veterinary application.

significant attention in recent years, founding the new interdisciplinary area called "network biology" (105). In network biology, biological systems are illustrated in the form of nodes and edges (105). Different types of networks such as signal transduction networks, protein-protein interaction networks, gene regulatory networks, and metabolic networks contain complex information about their relevant systems (104, 105). Nodes can represent genes, proteins, or metabolites, while edges represent interactions or relationships, according to the type of network (104). Network biology approaches are highly useful in investigations of hub nodes and drug targets, as well as identification of key components involved in regulating biological systems (47, 104, 106, 107).

Protein structure modeling, visualization, and validation

It was once challenging to predict a protein's 3D structure from its amino acid sequence. Now, these predictions are facilitated by improvements in protein structure prediction methods as well as the development of AlphaFold, a deep learning–based tool for protein structure modeling (108–110). When the target protein structure cannot be elucidated by experimental techniques, computational approaches become extremely important (110). These approaches can be used to predict protein structure, and the predicted structure can be utilized in drug screening. Additionally, computational approaches are used for predicting protein-protein interactions, structural comparison, and alignment (110, 111). In addition, several tools have been developed for visualization, refinement, and validation of the predicted 3D protein model. PyMOL is most often used tool for visualization, while Swiss PDB viewer, Rampage, PROCHECK, and Structural Analysis and Verification Server are extensively used for evaluation and model validation (109, 112, 113). With use of these tools, we can improve the quality of predicted models for further research (https://saves.mbi.ucla.edu/). Pan et al. predicted the cow milk 3D structures of αs1-CA and β-CA using I-TASSER to understand its dynamics (114). Additional research has modeled protein structure using computational approaches for livestock therapeutics development (115–117).

Binding site prediction

In drug discovery and design, binding site prediction is a crucial and significant step. A protein's 3D structure must be understood to identify amino acid residues present in the binding site. In order to learn more about the binding site and other sites, such as allosteric sites, computational tools are available to measure the area and volume of cavities in proteins (110). In vetinformatics, precise knowledge of the binding site is required to elucidate receptor–ligand interactions. Some molecular modeling and docking software packages offer the capability to predict and define the binding site prior to the docking simulation. Additionally, some web-based tools like CASTp and COACH are used for binding site predictions (118, 119).

Drug discovery and design for the management of livestock disease

The emergence of novel diseases decreases livestock productivity and represents a pressing challenge in the field of veterinary science. Effective treatments are unavailable for many diseases (6). Therefore, there is an urgent need to use vetinformatics approaches to identify novel lead molecules for drug development (5). In the process of developing new drugs, computational methods act as a valuable resource (110, 120). Finding a small molecule that can geometrically and chemically fit in a cavity of a macromolecular target is the aim of computer-assisted drug discovery programs (109, 110). Recent developments in computational approaches have facilitated the estimation of receptor–ligand binding energy through molecular docking simulations, prediction of pharmacokinetics and pharmacodynamics, and optimization of lead molecules (121). Due to advancements in computer power and algorithms, the field of drug discovery and design has achieved significant progress. For developing models and tools for drug discovery and design, computational methods including the hidden markov model, artificial neural networks, support vector machines, and genetic algorithms are frequently employed (109, 110, 121). In order to accelerate the drug development process, several issues have been solved, leading to a major improvement in these approaches and tools to reduce the time and cost of drug discovery programs (5, 30, 109). Several approaches and methods that play significant roles in veterinary drug discovery programs are highlighted in the following sections (Figure 2).

Molecular docking and virtual screening for identification of lead compounds

Recent developments in computational approaches have made it possible to predict molecular receptor–ligand interactions in the bound or complex state with perfect accuracy (110, 122, 123). To predict the interaction of small compounds with macromolecular targets, software such as AutoDock, AutoDock Vina, Glide, and Discovery studio are available (109). These programs can be used to screen a wide range of prospective compounds, look for new compounds with specific binding properties, or test available medicines with functional group alterations using molecular docking and virtual screening (109, 110, 122, 124). Recently, *in silico* studies predicted the lead compounds for drug development against porcine reproductive and respiratory syndrome virus (PRRSV) *via* the screening of 97,999 natural compounds from the ZINC database (6). The compounds 7-deacetyl-7-oxogedunin, kulactone, and nimocin were also identified as potential multi-target leads for the inhibition of porcine CD163 scavenger receptor cysteine-rich domain 5 (CD163-SRCR5), as well as non-structural protein 4 (Nsp4) and Nsp10 of PRRSV (5). The inhibitors of the imidazole glycerophosphate dehydratase protein in *Staphylococcus xylosus* were also identified through virtual screening (117).

ADMET and PAINs activity prediction of lead compounds

The primary criteria for sorting ligands in drug discovery programs involve its absorption, distribution, metabolism, excretion, and toxicity prediction, or ADMET (109, 121). These criteria act as a fundamental standard for testing candidate molecules. It is widely believed that every drug discovery program should consider these criteria, or Lipinski's rule of five, to evaluate orally active drugs (123, 125). In the early stages of the drug discovery process, abiding by these criteria is crucial for finding the most appropriate drug-like compounds, and it considerably reduces the late-stage failure of candidate molecules during preclinical or clinical trials (109, 110). Additionally, we can filter molecules that are related to pan-assay interference compounds (PAINS). Instead of directly affecting a specific target, PAINS typically respond non-specifically with many biological targets. In order to prevent non-specific binding and toxicity, a filter should be used (126, 127). Therefore, it is recognized as a cost-effective and time-saving approach in veterinary drug discovery program (5, 109, 110).

Pharmacophore and quantitative structure–activity relationship modeling

A pharmacophore is a molecular framework containing essential features of a drug's active component. Pharmacophore modeling is extensively used in the development of novel compounds (109, 110, 128). It can be used to represent and distinguish molecules at a 2D or 3D level by schematically illustrating the essential components of molecular recognition (110, 123). Relatedly, quantitative structure–activity relationship modeling is a widely used drug discovery approach that utilizes a molecule's physicochemical properties to predict its biological activity (110, 129). Both of these approaches can be used to find novel treatments for livestock diseases (5, 30).

Molecular dynamics simulation of proteins and protein–ligand complexes to determine their dynamics and behavior during interactions

Molecular dynamics simulation is used to computationally visualize the movement and behavior of a molecular system at the atomic level (110, 130). It offers a wealth of knowledge regarding the interactions between proteins and ligands and provides complex structural information on macromolecular structures (109, 110). This knowledge is crucial for understanding the structure–function relationship among the target and its dynamics during protein–ligand interactions, ultimately supporting drug discovery processes (109). As a result, it is widely utilized to characterize protein–ligand interactions in modern drug discovery programs (6). Additionally, it is used to validate predicted protein models, understand the dynamics of protein folding and unfolding and protein–ligand dynamics, examine the effects of mutations on structures, and understand binding dynamics at other sites (5). A recent study described the role of DGAT1 missense non-synonymous single nucleotide polymorphisms (SNPs) in dairy cattle using computational approaches. The DGAT1 variants (W128R, W214R, C215G, P245R, and W459G) were analyzed initially through sequence- and structure-based tools, then evaluated using molecular dynamics simulation to understand their structural and conformational dynamics compared to wild-type structures and improve milk quality in cattle (47).

Designing vaccines for livestock diseases

Emerging pathogens are a major threat to livestock productivity that requires the identification of vaccine candidates in order to ensure long-term protection of animals (33, 131). In order to provide broad-spectrum and long-term protection against different viral and bacterial diseases, new approaches to vaccine development must be created (10, 132). In the post-genome era, identifying specific antigenic regions to activate certain arms of the immune system was a major challenge (115, 133). To address this issue, computational vaccine design has been a major area of interest for researchers over the last two decades. Several tools and web-based resources have been developed that have proven useful in vaccine design (133, 134). Researchers can now utilize advanced vetinformatics approaches to design vaccines that provide protection against livestock diseases (33, 115, 131).

Machine and deep learning approaches in livestock research

Machine and deep learning approaches have received significant attention from veterinary scientists (135, 136). Computers are equipped with an adaptive mechanism that enables them to learn from examples and experiences (137). Machine and deep learning provide information-processing capabilities for handling various types of real-life information (137). In order to make predictions or conclusions about target datasets, these algorithms often build mathematical models using sample datasets, also referred to as training datasets (137, 138). The recent advancements in artificial intelligence have made it even easier to analyze animal behavior in videos using machine vision and machine learning (139). The development of predictive models such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) based on modern machine techniques are helpful in livestock research (139, 140). It was shown that the development of a recurrent neural network (RNN) model with an LSTM could classify cattle behavior in a reasonable manner (141). Recently, CNN and Bidirectional Long Short-Term Memory (BiLSTM) were used for video-based identification of individual cattle (140), and C3D-ConvLSTM (Convolutional 3D- Convolutional Long Short-Term Memory) based model was used for cow behavior classification over 86% accuracy (142).

Enabled by advances in omics, an enormous amount of biological data is produced every day, and these large data sets allow researchers to build machine learning models in order to make relevant predictions and minimize the expense and duration of experiments (137, 138, 143). These approaches play important roles in different areas of vetinformatics, such as gene discovery and genome annotation, gene expression analysis,

TABLE 1 List of important databases for research in vetinformatics.

S. No.	Database	Application	Availability	References
1	National Center for Biotechnology Information (NCBI)	Offers resources for research and development in different areas of the life sciences, including veterinary and animal sciences	https://www.ncbi.nlm.nih.gov/	(144)
2	Uniprot	Provides comprehensive resources related to protein sequences, including relevant functional and structural information	https://www.uniprot.org/	(145)
3	Pfam	Database of protein families used for domain analysis and related information	https://pfam.xfam.org/	(146)
4	Protein Data Bank (PDB)	Structural database with information on macromolecules' three-dimensional structures, which is useful in drug discovery and structural bioinformatics	https://www.rcsb.org/	(147)
5	AlphaFold Protein Structure Database	Database containing predicted 3D structures of human proteins and other key proteins	https://alphafold.ebi.ac.uk/	(148)
6	PubChem	NCBI Database of small molecules and related information, including the structure of chemical compounds, applicable for use in molecular docking and virtual screening	https://pubchem.ncbi.nlm.nih.gov/	(35)
7	ZINC	Database of commercially available molecules for use in virtual screening	https://zinc.docking.org/	(149)
8	GEO	Functional genomics database hosted at NCBI offering gene expression profiles that have been provided by an international scientific community	https://www.ncbi.nlm.nih.gov/geo/	(150)
9	Sequence Read Archive (SRA)	The largest collection of publicly accessible high-throughput sequencing data, comprising NGS data provided by the international scientific community for use in research and integration of multi-omics data	https://www.ncbi.nlm.nih.gov/sra	(151)
10	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Database containing pathways for understanding biological systems	https://www.genome.jp/kegg/	(152)

(Continued)

TABLE 1 (Continued)

S. No.	Database	Application	Availability	References
11	Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	Database containing information about protein–protein interactions derived from experimental, computational, and text-mining techniques	https://string-db.org/	(153)
12	BioModels	Collection of mathematical models in standard file formats for further analysis and integration of biological systems	https://www.ebi.ac.uk/biomodels/	(154)
13	Bovine Genome Database	Database providing genomics resources and tools for bovine research	https://bovinegenome.elsiklab.missouri.edu/#:$\sim$:text=The%20Bovine%20Genome%20Database%20supportsthereford%20cow%2C%20L1%20Dominette%2001449	(37)
14	Porcine Translational Research Database	Database providing genomic and proteomic information involving pigs	https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/diet-genomics-and-immunology-laboratory/docs/dgil-porcine-translational-research-database/	(36)
15	Animal-ImputeDB	Database and resource for the study of animal genotype imputation	http://gong_lab.hzau.edu.cn/Animal_ImputeDB/#!/	(38)
16	SNPRBb	Database containing trait-specific SNP resources for *Bubalus bubalis*, including information on important genomic variants	http://cabgrid.res.in:8080/snprbb/home.php	(155)
17	BuffSatDb	Water buffalo (*Bubalus bubalis*) genome-wide microsatellite database	http://webapp.cabgrid.res.in/buffsatdb/index.html	(156)
18	National Animal Genome Research Program	Comprehensive resource for research in livestock genomics	https://www.animalgenome.org/	(4)
19	Chickspress	Gene expression database for chicken tissues	https://geneatlas.arl.arizona.edu/	(157)
20	Ensembl genome browser	Genome browser containing genomic information of several livestock animals	https://www.ensembl.org/index.html	(158)

TABLE 2 A list of popular computational software available for livestock research.

S. No.	Software	Application	Availability	References
1	Basic Local Alignment Search Tool (BLAST)	Finds homologous and paralogous sequences and provides similarity searching	https://blast.ncbi.nlm.nih.gov/Blast.cgi	(72)
2	SRA Toolkit	Creating FASTQ files from SRA	https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit	(162, 163)
3	FastQC	Assesses the quality of raw sequencing data produced by NGS platforms	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	(164)
4	Trimmomatic	Trims reads for Illumina NGS data	http://www.usadellab.org/cms/?page=trimmomatic	(165)
5	Cutadapt	Identifies primers, adapter sequences, poly-A tails, and other regions, and removes them from sequencing reads	https://cutadapt.readthedocs.io/en/stable/	(166)
6	fastp	Preprocessing of FASTQ files which includes quality control, adapter trimming, quality filtering etc	https://github.com/OpenGene/fastp	(167)
7	HISAT2	Maps next-generation sequencing reads quickly and accurately	http://daehwankimlab.github.io/hisat2/	(168)
8	Samtools	Used for post-processing of short DNA sequence read alignments	http://www.htslib.org/	(169)
9	Bowtie 2	Aligns sequencing reads to reference sequences	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	(170)
10	BWA	Mapping sequence reads to reference genome	https://bio-bwa.sourceforge.net/	(171)
11	Trinity	Assembles transcriptome or RNA-Seq data produced by the Illumina NGS platform using the *de novo* approach	https://github.com/trinityrnaseq/trinityrnaseq/wiki	(172)
12	edgeR	R package used to identify differentially expressed genes using RNA-Seq data	https://bioconductor.org/packages/release/bioc/html/edgeR.html	(173)

(Continued)

TABLE 2 (Continued)

S. No.	Software	Application	Availability	References
13	DESeq2	Differential gene expression analysis	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	(174)
14	WGCNA	Co-expression network analysis	https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/	(175)
15	GATK	Identification of variants using high-throughput sequencing data	yandell-lab.org/software/maker.html	(176)
16	Molecular Evolutionary Genetics Analysis (MEGA)	Creates phylogenetic trees and performs statistical analyses of molecular evolution	https://www.megasoftware.net/	(77)
17	Velvet	Handles *de novo* genome assembly using short-read sequencing data	https://www.ebi.ac.uk/$\sim$zerbino/velvet/	(177)
18	SPAdes	Single-cell and multi-cell genome assembly	https://cab.spbu.ru/software/spades/	(52)
19	MAKER	Genome annotation	https://github.com/Yandell-Lab/maker	(53)
20	REVIGO	Summarizes and visually represents gene ontology terms	http://revigo.irb.hr/	(178)
21	Multi-Experiment Viewer (WebMeV)	Creates analyses and visualizations of genomic data	https://webmev.tm4.org/#/about	(179)
22	Gene Set Enrichment Analysis (GSEA)	Facilitates the analysis and interpretation of gene expression data	https://www.gsea-msigdb.org/gsea/index.jsp	(180)
23	DIAMOND	Performs comparatively rapid sequence alignment of proteins or translated DNA sequences in order to examine of large amounts of sequence data	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/diamond/	(181)
24	Blast2GO	Used to perform genomic data annotation and gene ontology analysis	https://www.blast2go.com/	(182)
25	Cytoscape	Offers tools and plugins for visualization and research in network science and network biology	https://cytoscape.org/	(183)

(Continued)

TABLE 2 (Continued)

S. No.	Software	Application	Availability	References
26	AlphaFold 2	Uses deep learning methods to predict protein structure using amino acid sequences	https://github.com/deepmind/alphafold	(108)
27	PyMOL	Offers tools for the visualization and analysis of macromolecular structures in 3D	https://pymol.org/2/	(112)
28	Swiss PDB Viewer	Enables simultaneous analysis of protein structures, including calculation of H-bonds, angles, and atom distances as well as comparison and alignment of macromolecular structures	https://spdbv.vital-it.ch/	(113)
29	Chimera	Offers tools for visualizing and analyzing molecular structures and creating density maps, motions, and sequence and structural alignments, producing high-quality images	https://www.cgl.ucsf.edu/chimera/	(184)
30	Protein Variation Effect Analyzer (PROVEAN)	Predicts how an amino acid substitution or indel may affect the biological function of a protein	http://provean.jcvi.org/index.php	(185)
31	MarvinSketch	Offers tools for the conversion of structural file formats as well as for drawing, editing, importing, and exporting chemical structures and calculating their properties	https://chemaxon.com/products/marvin	(186)
32	CASTp	Binding site prediction	http://sts.bioe.uic.edu/castp/index.html?2was	(119)
33	AutoDock	Offers tools for molecular docking studies	http://autodock.scripps.edu/	(124)
34	SwissADME	Physicochemical properties, Pharmacokinetics, Druglikeness prediction	http://www.swissadme.ch/	(187)
35	GROningen MAchine for Chemical Simulations (GROMACS)	Offers high-performance molecular dynamics tools for simulations of proteins, lipids, and nucleic acids	http://www.gromacs.org/	(188)

drug target prediction, protein modeling, drug discovery, text mining, digital image processing, and helpful in precision livestock farming (137, 138, 143).

Development of databases and tools for vetinformatics

Databases and tools related to veterinary science are essential for computer-based examinations of livestock data (30, 32, 33). Several databases and tools are available, but most databases contain information about many organisms (10) (Table 1). Due to recent developments in the area of vetinformatics, some animal-specific databases have been developed in recent years, but their availability is still insufficient (36–38). In the post-genomic era, large multi-omics data sets about livestock animals are urgently needed to develop species-specific databases to support veterinary science. Species-specific databases would help veterinary researchers easily find information about target animals. In addition, the availability of multi-omics data will help to improve the prediction, development, and accuracy of new algorithms that solve problems related to animal breeding, develop disease diagnostics, and offer solutions that increase livestock productivity and sustainability (138, 159–161). Some of the important software used for livestock research is highlighted in Table 2.

Future perspectives on vetinformatics

Since the beginning of the human genome project, the use of computers in biology has drawn significant interest and it is currently an essential tool in biological research. In the twenty-first century, it is difficult to imagine a novel discovery that does not rely on computational methods. Because computer software has been involved in most biological studies worldwide in the current omics era, many top research groups believe that integration of computers with biology has immense potential to decode complex biological systems, enabling the discovery of novel therapeutics and other useful information for the betterment of society. Therefore, vetinformatics will eventually become a crucial component of every veterinary science research lab. The management of big data in biology and veterinary science will also demand vetinformatics experts, who will reduce experimental work load and expense. As the human population grows, requiring commensurate increases in food production, it will be necessary to increase livestock productivity, advance animal breeding programs, improve the nutritional quality of animal products, and develop disease prevention and management strategies for animal welfare. This can be accomplished with the help of vetinformatics approaches that visualize the complexity of livestock systems in order to design solutions that meet our demands for higher livestock productivity.

Conclusion

In recent years, vetinformatics has emerged as a vital subject and a popular interdisciplinary research area in veterinary sciences. The strength of vetinformatics and the ability of its methods to tackle challenging projects in veterinary sciences were highlighted in this review. Databases and other tools available for livestock research, along with their applications and availability, were also included. Vetinformatics approaches have proven their ability to resolve a variety of problems in veterinary science. To develop vetinformatics tools and databases that successfully target livestock systems for quality veterinary services, more resources need to be developed. Therefore, a conversation is needed in the veterinary science community that encourages the implementation of vetinformatics to understand livestock systems for the enhancement of animal welfare and drug discovery.

Author contributions

J-MK and RKP developed the idea for this review article and its coverage. RKP wrote the manuscript. J-MK supervised the work and edited the manuscript. Both authors have read the final manuscript and approved the submission.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2022R1A2C1005830).

Acknowledgments

The authors thank Chung-Ang University, Anseong-si, Republic of Korea for providing high-performance computing and other necessary facilities.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Grandin T. Introduction: the contribution of animals to human welfare. *Rev Sci Tech.* (2018) 37:15–35. doi: 10.20506/rst.37.1.2737

2. Randolph TF, Schelling E, Grace D, Nicholson CF, Leroy JL, Cole DC, et al. Invited review: role of livestock in human nutrition and health for poverty reduction in developing countries. *J Anim Sci.* (2007) 85:2788–800. doi: 10.2527/jas.2007-0467

3. Dupjan S, Dawkins MS. Animal welfare and resistance to disease: interaction of affective states and the immune system. *Front Vet Sci.* (2022) 9:929805. doi: 10.3389/fvets.2022.929805

4. Rexroad C, Vallet J, Matukumalli LK, Reecy J, Bickhart D, Blackburn H, et al. Genome to phenome: improving animal health, production, and well-being - a new usda blueprint for animal genome research 2018-2027. *Front Genet.* (2019) 10:327. doi: 10.3389/fgene.2019.00327

5. Pathak RK, Kim D-Y, Lim B, Kim J-M. Investigating multi-target antiviral compounds by screening of phytochemicals from neem (*Azadirachta indica*) against PRRSV: a vetinformatics approach. *Front Vet Sci.* (2022) 9:854528. doi: 10.3389/fvets.2022.854528

6. Pathak RK, Seo YJ, Kim JM. Structural insights into inhibition of Prrsv Nsp4 revealed by structure-based virtual screening, molecular dynamics, and Mm-Pbsa studies. *J Biol Eng.* (2022) 16:4. doi: 10.1186/s13036-022-00284-x

7. Kim DY, Kim JM. Multi-Omics integration strategies for animal epigenetic studies - a review. *Anim Biosci.* (2021) 34:1271–82. doi: 10.5713/ab.21.0042

8. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinform.* (2019) 20:1981–96. doi: 10.1093/bib/bby063

9. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol.* (2011) 7:e1002021. doi: 10.1371/journal.pcbi.1002021

10. Pathak RK, Singh DB, Singh R. Introduction to basics of bioinformatics. In: Singh DB, Pathak RK, editors. *Bioinformatics*. Academic Press Elsevier (2022). p. 1–15. doi: 10.1016/B978-0-323-89775-4.00006-7

11. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med.* (1944) 79:137–58.

12. Griffiths JF, Griffiths AJ, Wessler SR, Lewontin RC, Gelbart WM, Suzuki DT, et al. *An Introduction to Genetic Analysis*. Macmillan (2005). Available online at: https://store.macmillanlearning.com/us/product/Introduction-to-Genetic-Analysis/p/1319114784

13. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol.* (1952) 36:39–56. doi: 10.1085/jgp.36.1.39

14. Tamm C, Shapiro HS, Lipshitz R, Chargaff E. Distribution density of nucleotides within a desoxyribonucleic acid chain. *J Biol Chem.* (1953) 203:673–88. doi: 10.1016/S0021-9258(19)52337-7

15. Watson JD, Crick FH. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature.* (1974) 248:765. doi: 10.1038/248765a0

16. Nirenberg M, Leder P. Rna codewords and protein synthesis: the effect of trinucleotides upon the binding of srna to ribosomes. *Science.* (1964) 145:1399–407. doi: 10.1126/science.145.3639.1399

17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA.* (1977) 74:5463–7. doi: 10.1073/pnas.74.12.5463

18. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA.* (1977) 74:560–4. doi: 10.1073/pnas.74.2.560

19. Jaskolski M, Dauter Z, Wlodawer A. A brief history of macromolecular crystallography, illustrated by a family tree and its nobel fruits. *FEBS J.* (2014) 281:3985–4009. doi: 10.1111/febs.12796

20. Sanger F, Thompson EO. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J.* (1953) 53:353–66. doi: 10.1042/bj0530353

21. Hagen JB. The origins of bioinformatics. *Nat Rev Genet.* (2000) 1:231–6. doi: 10.1038/35042090

22. Moody G. *Digital Code of Life : How Bioinformatics Is Revolutionizing Science, Medicine, and Business*. Hoboken, NJ: Wiley (2004) 389 p.

23. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* (1970) 48:443–53. doi: 10.1016/0022-2836(70)90057-4

24. Higgins DG, Sharp PM. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene.* (1988) 73:237–44. doi: 10.1016/0378-1119(88)90330-7

25. Ranwez V, Chantret N. *Strengths and Limits of Multiple Sequence Alignment and Filtering Methods* (2020). Available online at: https://hal.archives-ouvertes.fr/hal-02535389/document

26. Dayhoff M, Schwartz R, Orcutt B. A model of evolutionary change in proteins. In: Dayhoff M, editor. *Atlas of Protein Sequence and Structure* (1972). Available online at: https://chagall.med.cornell.edu/BioinfoCourse/PDFs/Lecture2/Dayhoff1978.pdf

27. Chial H. DNA sequencing technologies key to the human genome project. *Nat Educ.* (2008) 1:219. Available online at: https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/

28. Hood L, Rowen L. The human genome project: big science transforms biology and medicine. *Genome Med.* (2013) 5:79. doi: 10.1186/gm483

29. Coordinators NR. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* (2015) 43:D6–17. doi: 10.1093/nar/gku1130

30. Sujatha P, Kumarasamy P, Preetha S, Balachandran P. Vetinformatics: a new paradigm for quality veterinary services. *Res Rev J Vet Sci Technol.* (2018) 5:16–9. doi: 10.37591/rrjovst.v5i2.537

31. Tiwary BK. Farm animal informatics. In: Tiwary BK, editor. *Bioinformatics and Computational Biology*. Singapore: Springer (2022). p. 203–18. doi: 10.1007/978-981-16-4241-8_11

32. Hardy T. Animal bioinformatics. *EJBI.* (2021) 17:9–10. Available online at: https://www.ejbi.org/scholarly-articles/animal-bioinformatics-8691.html

33. Kaikabo A, Kalshingi H. Concepts of bioinformatics and its application in veterinary research and vaccines development. *Nigerian Vet J.* (2007) 28:39–46. doi: 10.4314/nvj.v28i2.3554

34. Byrne C, Logas J. The future of technology and computers in veterinary medicine. *Diagnost Ther Vet Dermatol.* (2021) 245–50. doi: 10.1002/9781119680642.ch26

35. Hahnke VD, Kim S, Bolton EE. Pubchem chemical structure standardization. *J Cheminform.* (2018) 10:36. doi: 10.1186/s13321-018-0293-8

36. Dawson HD, Chen C, Gaynor B, Shao J, Urban JF Jr. The porcine translational research database: a manually curated, genomics and proteomics-based research resource. *BMC Genomics.* (2017) 18:643. doi: 10.1186/s12864-017-4009-7

37. Shamimuzzaman M, Le Tourneau JJ, Unni DR, Diesh CM, Triant DA, Walsh AT, et al. Bovine genome database: new annotation tools for a new reference genome. *Nucleic Acids Res.* (2020) 48:D676–81. doi: 10.1093/nar/gkz944

38. Yang W, Yang Y, Zhao C, Yang K, Wang D, Yang J, et al. Animal-Imputedb: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Res.* (2020) 48:D659–67. doi: 10.1093/nar/gkz854

39. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp.* (2012) 2:3. doi: 10.1186/2042-5783-2-3

40. Guillemin N, Horvatic A, Kules J, Galan A, Mrljak V, Bhide M. Omics approaches to probe markers of disease resistance in animal sciences. *Mol Biosyst.* (2016) 12:2036–46. doi: 10.1039/C6MB00220J

41. Ghosh S, Dasgupta R. Machine learning in the study of animal health and veterinary sciences. In: Ghosh S, Dasgupta R, editors. *Machine Learning in Biological Sciences*. Singapore: Springer (2022). p. 251–9. doi: 10.1007/978-981-16-8881-2_29

42. Ezanno P, Picault S, Beaunee G, Bailly X, Munoz F, Duboz R, et al. Research perspectives on animal health in the era of artificial intelligence. *Vet Res.* (2021) 52:40. doi: 10.1186/s13567-021-00902-4

43. Morrison-Smith S, Boucher C, Sarcevic A, Noyes N, O'Brien C, Cuadros N, et al. Challenges in large-scale bioinformatics projects. *Hum Soc Sci Commun.* (2022) 9:125. doi: 10.1057/s41599-022-01141-4

44. Soetan KO, Awosanya EA. Bioinformatics and its application in animal health: a review. *Trop Vet.* (2015) 33:3–22. Available online at: https://www.ajol.info/index.php/tv/article/view/160158

45. Li Q, Freeman LM, Rush JE, Huggins GS, Kennedy AD, Labuda JA, et al. Veterinary medicine and multi-omics research for future nutrition targets: metabolomics and transcriptomics of the common degenerative mitral valve disease in dogs. *Omics.* (2015) 19:461–70. doi: 10.1089/omi.2015.0057

46. Jianghong W, Li X, Xu X. Multi-Omics approaches to study complex traits in domestic animals. *Front Syst Biol.* (2021) 1:771644. doi: 10.3389/fsysb.2021.771644

47. Pathak RK, Lim B, Park Y, Kim JM. Unraveling structural and conformational dynamics of Dgat1 missense Nssnps in dairy cattle. *Sci Rep.* (2022) 12:4873. doi: 10.1038/s41598-022-08833-6

48. Park Y, Park YB, Lim SW, Lim B, Kim JM. Time series ovarian transcriptome analyses of the porcine estrous cycle reveals gene expression changes during steroid metabolism and corpus luteum development. *Animals.* (2022) 12:376. doi: 10.3390/ani12030376

49. Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: where are we now? *Proc Natl Acad Sci USA.* (2021) 118:e2109019118. doi: 10.1073/pnas.2109019118

50. Baker M. *De novo* genome assembly: what every biologist should know. *Nat Methods.* (2012) 9:333–7. doi: 10.1038/nmeth.1935

51. Thrash A, Hoffmann F, Perkins A. Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics.* (2020) 21:249. doi: 10.1186/s12859-020-3382-4

52. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* (2012) 19:455–77. doi: 10.1089/cmb.2012.0021

53. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* (2008) 18:188–96. doi: 10.1101/gr.6743907

54. Taylor JF, Whitacre LK, Hoff JL, Tizioto PC, Kim J, Decker JE, et al. Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. *Genet Sel Evol.* (2016) 48:59. doi: 10.1186/s12711-016-0237-6

55. Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. *Nat Commun.* (2022) 13:910. doi: 10.1038/s41467-022-28605-0

56. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, bos taurus. *Genome Biol.* (2009) 10:R42. doi: 10.1186/gb-2009-10-4-r42

57. Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun.* (2021) 12:1821. doi: 10.1038/s41467-021-22100-8

58. Bovo S, Schiavo G, Bolner M, Ballan M, Fontanesi L. Mining livestock genome datasets for an unconventional characterization of animal DNA viromes. *Genomics.* (2022) 114:110312. doi: 10.1016/j.ygeno.2022.110312

59. Xia XH. Bioinformatics and drug discovery. *Curr Top Med Chem.* (2017) 17:1709–26. doi: 10.2174/1568026617666161116143440

60. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. *De novo* assembly of the cattle reference genome with single-molecule sequencing. *Gigascience.* (2020) 9:giaa021. doi: 10.1093/gigascience/giaa021

61. Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, et al. Comparison of Rna-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Front Genet.* (2018) 9:636. doi: 10.3389/fgene.2018.00636

62. Wolf JB. Principles of transcriptome analysis and gene expression quantification: an Rna-Seq tutorial. *Mol Ecol Resour.* (2013) 13:559–72. doi: 10.1111/1755-0998.12109

63. Jang MJ, Lim C, Lim B, Kim JM. Integrated multiple transcriptomes in oviductal tissue across the porcine estrous cycle reveal functional roles in oocyte maturation and transport. *J Anim Sci.* (2022) 100:skab364. doi: 10.1093/jas/skab364

64. Lam JKW, Chow MYT, Zhang Y, Leung SWS. Sirna versus mirna as therapeutics for gene silencing. *Mol Ther-Nucl Acids.* (2015) 4:e252. doi: 10.1038/mtna.2015.23

65. Do DN, Dudemaine PL, Mathur M, Suravajhala P, Zhao X, Ibeagha-Awemu EM. Mirna regulatory functions in farm animal diseases, and biomarker potentials for effective therapies. *Int J Mol Sci.* (2021) 22:3080. doi: 10.3390/ijms22063080

66. Miretti S, Lecchi C, Ceciliani F, Baratta M. Micrornas as biomarkers for animal health and welfare in livestock. *Front Vet Sci.* (2020) 7:578193. doi: 10.3389/fvets.2020.578193

67. Li Z, Wang H, Chen L, Wang L, Liu X, Ru C, et al. Identification and characterization of novel and differentially expressed micro Rna S in peripheral blood from healthy and mastitis holstein cattle by deep sequencing. *Anim Genet.* (2014) 45:20–7. doi: 10.1111/age.12096

68. Chen Z, Zhou J, Wang X, Zhang Y, Lu X, Fan Y, et al. Screening candidate micror-15a-Irak2 regulatory pairs for predicting the response to staphylococcus aureus-induced mastitis in dairy cows. *J Dairy Res.* (2019) 86:425–31. doi: 10.1017/S0022029919000785

69. Dana H, Chalbatani GM, Mahmoodzadeh H, Karimloo R, Rezaiean O, Moradzadeh A, et al. Molecular mechanisms and biological functions of sirna. *Int J Biomed Sci.* (2017) 13:48–57. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5542916/

70. Kumar H, Jang YN, Kim K, Park J, Jung MW, Park JE. Compositional and functional characteristics of swine slurry microbes through 16s Rrna metagenomic sequencing approach. *Animals.* (2020) 10:1372. doi: 10.3390/ani10081372

71. de Menezes AB, Lewis E, O'Donovan M, O'Neill BF, Clipson N, Doyle EM. Microbiome analysis of dairy cows fed pasture or total mixed ration diets. *FEMS Microbiol Ecol.* (2011) 78:256–65. doi: 10.1111/j.1574-6941.2011.01151.x

72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* (1990) 215:403–10. doi: 10.1016/S0022-2836(05)80360-2

73. Ajayi OO, Peters SO, De Donato M, Sowande SO, Mujibi FDN, Morenikeji OB, et al. Computational genome-wide identification of heat shock protein genes in the bovine genome. *F1000Res.* (2018) 7:1504. doi: 10.12688/f1000research.16058.1

74. Samuel B, Dinka H. *In silico* analysis of the promoter region of olfactory receptors in cattle (*Bos indicus*) to understand its gene regulation. *Nucleosides Nucleotides Nucleic Acids.* (2020) 39:853–65. doi: 10.1080/15257770.2020.1711524

75. Quan JQ, Cai Y, Yang TL, Ge QY, Jiao T, Zhao SG. Phylogeny and conservation priority assessment of asian domestic chicken genetic resources. *Glob Ecol Conserv.* (2020) 22:e00944. doi: 10.1016/j.gecco.2020.e00944

76. Olvera A, Busquets N, Cortey M, de Deus N, Ganges L, Nunez JI, et al. Applying phylogenetic analysis to viral livestock diseases: moving beyond molecular typing. *Vet J.* (2010) 184:130–7. doi: 10.1016/j.tvjl.2009.02.015

77. Tamura K, Stecher G, Kumar S. Mega11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol.* (2021) 38:3022–7. doi: 10.1093/molbev/msab120

78. Retief JD. Phylogenetic analysis using phylip. In: Misener S, Krawetz SA, editors. *Bioinformatics Methods and Protocols*. Humana Totowa, NJ: Springer (2000). p. 243–58. doi: 10.1385/1-59259-192-2:243

79. Soltis PS, Soltis DE. Applying the bootstrap in phylogeny reconstruction. *Stat Sci.* (2003) 18:256–67. doi: 10.1214/ss/1063994980

80. Letunic I, Bork P. Interactive tree of life (Itol) V5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* (2021) 49:W293–6. doi: 10.1093/nar/gkab301

81. Page RD. Tree view: an application to display phylogenetic trees on personal computers. *Bioinformatics.* (1996) 12:357–8. doi: 10.1093/bioinformatics/12.4.357

82. Deng Y, Guan SH, Wang S, Hao G, Rasmussen TB. The detection and phylogenetic analysis of bovine hepacivirus in China. *Biomed Res Int.* (2018) 2018:6216853. doi: 10.1155/2018/6216853

83. Singhla T, Boonsri K, Kreausukon K, Modethed W, Pringproa K, Sthitmatee N, et al. Molecular characterization and phylogenetic analysis of lumpy skin disease virus collected from outbreaks in northern Thailand in 2021. *Vet Sci.* (2022) 9:194. doi: 10.3390/vetsci9040194

84. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-Wide association studies. *Nat Rev Methods Prim.* (2021) 1:1–21. doi: 10.1038/s43586-021-00056-9

85. Cheng J, Fernando R, Cheng H, Kachman SD, Lim K, Harding JCS, et al. Genome-Wide association study of disease resilience traits from a natural polymicrobial disease challenge model in pigs identifies the importance of the major histocompatibility complex region. *G3.* (2022) 12:jkab441. doi: 10.1093/g3journal/jkab441

86. Mkize N, Maiwashe A, Dzama K, Dube B, Mapholi N. Suitability of gwas as a tool to discover Snps associated with tick resistance in cattle: a review. *Pathogens.* (2021) 10:1604. doi: 10.3390/pathogens10121604

87. Chu BB, Keys KL, German CA, Zhou H, Zhou JJ, Sobel EM, et al. Iterative hard thresholding in genome-wide association studies: generalized linear models, prior weights, and double sparsity. *Gigascience.* (2020) 9:giaa044. doi: 10.1093/gigascience/giaa044

88. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* (2010) 42:355–60. doi: 10.1038/ng.546

89. Shenstone E, Cooper J, Rice B, Bohn M, Jamann TM, Lipka AE. An assessment of the performance of the logistic mixed model for analyzing binary traits in maize and sorghum diversity panels. *PLoS ONE.* (2018) 13:e0207752. doi: 10.1371/journal.pone.0207752

90. Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep.* (2016) 6:19444. doi: 10.1038/srep19 444

91. Zhang YW, Tamba CL, Wen YJ, Li P, Ren WL, Ni YL, et al. Mrmlm V4.0.2: an R platform for multi-locus genome-wide association studies. *Genomics Proteomics Bioinformatics.* (2020) 18:481–7. doi: 10.1016/j.gpb.2020.06. 006

92. Wen YJ, Zhang YW, Zhang J, Feng JY, Zhang YM. The improved fastmremma and Gcim algorithms for genome-wide association and linkage studies in large mapping populations. *Crop J.* (2020) 8:723–32. doi: 10.1016/j.cj.2020.04.008

93. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* (2007) 81:559–75. doi: 10.1086/519795

94. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. Genabel: an R library for genome-wide association analysis. *Bioinformatics.* (2007) 23:1294–6. doi: 10.1093/bioinformatics/btm108

95. Curtis RE, Kinnaird P, Xing EP, editors. Genamap: visualization strategies for structured association mapping. In: *2011 IEEE Symposium on Biological Data Visualization (BioVis).* Providence, RI: IEEE (2011). doi: 10.1109/BioVis.2011.6094052

96. Zhou X, Stephens M. Genome-Wide efficient mixed-model analysis for association studies. *Nat Genet.* (2012) 44:821–4. doi: 10.1038/ng.2310

97. Liu Z, Li H, Zhong Z, Jiang S. A whole genome sequencing-based genome-wide association study reveals the potential associations of teat number in qingping pigs. *Animals.* (2022) 12:1057. doi: 10.3390/ani12091057

98. Uemoto Y, Ichinoseki K, Matsumoto T, Oka N, Takamori H, Kadowaki H, et al. Genome-Wide association studies for production, respiratory disease, and immune-related traits in landrace pigs. *Sci Rep.* (2021) 11:15823. doi: 10.1038/s41598-021-95339-2

99. Kim S, Lim B, Cho J, Lee S, Dang CG, Jeon JH, et al. Genome-Wide identification of candidate genes for milk production traits in korean holstein cattle. *Animals.* (2021) 11:1392. doi: 10.3390/ani11051392

100. Lee J, Lee SM, Lim B, Park J, Song KL, Jeon JH, et al. Estimation of variance components and genomic prediction for individual birth weight using three different genome-wide snp platforms in yorkshire pigs. *Animals.* (2020) 10:2219. doi: 10.3390/ani10122219

101. Adhil M, Agarwal M, Achutharao P, Talukder AK. Advanced computational methods, ngs tools, and software for mammalian systems biology. In: Kadarmideen HN, editor. *Systems Biology in Animal Production and Health, Vol. 1.* Springer (2016). p. 117–51. doi: 10.1007/978-3-319-43335-6_6

102. Headon D. Systems Biology And Livestock Production. *Animal.* (2013) 7:1959–63. doi: 10.1017/S1751731113000980

103. Kadarmideen HN. Genomics to systems biology in animal and veterinary sciences: progress, lessons and opportunities. *Livest Sci.* (2014) 166:232–48. doi: 10.1016/j.livsci.2014.04.028

104. Pathak RK, Singh DB. Systems biology approaches for food and health. In: Sharma TR, Deshmukh R, Sonah H, editors. *Advances in Agri-Food Biotechnology.* Singapore: Springer Nature (2020). p. 409–26. doi: 10.1007/978-981-15-2874-3_16

105. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* (2004) 5:101–13. doi: 10.1038/nrg1272

106. Ma J, Wang J, Ghoraie LS, Men X, Liu LN, Dai PG. Network-Based method for drug target discovery at the isoform level. *Sci Rep.* (2019) 9:13868. doi: 10.1038/s41598-019-50224-x

107. Lim D, Kim N-K, Park H-S, Lee S-H, Cho Y-M, Oh SJ, et al. Identification of candidate genes related to bovine marbling using protein-protein interaction networks. *Int J Biol Sci.* (2011) 7:992. doi: 10.7150/ijbs.7.992

108. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. *Nature.* (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

109. Pathak RK, Singh DB, Sagar M, Baunthiyal M, Kumar A. Computational approaches in drug discovery and design. In: Singh DB, editor. *Computer-Aided Drug Design.* Singapore: Springer (2020). p. 1–21. doi: 10.1007/978-981-15-6815-2_1

110. Singh DB, Pathak RK. Computational approaches in drug designing and their applications. In: Gupta N, Gupta V, editors. *Experimental Protocols in Biotechnology.* Humana New York, NY: Springer (2020). p. 95–117. doi: 10.1007/978-1-0716-0607-0_6

111. Vakser IA. Protein-Protein docking: from interaction to interactome. *Biophys J.* (2014) 107:1785–93. doi: 10.1016/j.bpj.2014.08.033

112. DeLano WL. Pymol: an open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr.* (2002) 40:82–92. Available online at: https://legacy.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf

113. Johansson MU, Zoete V, Michielin O, Guex N. Defining and searching for structural motifs using deepview/Swiss-Pdbviewer. *BMC Bioinformatics.* (2012) 13:173. doi: 10.1186/1471-2105-13-173

114. Pan F, Li JX, Zhao L, Tuersuntuoheti T, Mehmood A, Zhou N, et al. A Molecular docking and molecular dynamics simulation study on the interaction between cyanidin-3-O-glucoside and major proteins in cow's milk. *J Food Biochem.* (2021) 45:e13570. doi: 10.1111/jfbc.13570

115. Muguntha SP, Mani Chandra H. A computational reverse vaccinology approach for the design and development of multi-epitopic vaccine against avian pathogen mycoplasma gallisepticum. *Front Vet Sci.* (2021) 8:721061. doi: 10.3389/fvets.2021.721061

116. Thakuria D, Khangembam VC, Pant V, Bhat RAH, Tandel RS, C S, et al. Anti-Oomycete activity of chlorhexidine gluconate: molecular docking and *in vitro* studies. *Front Vet Sci.* (2022) 9:909570. doi: 10.3389/fvets.2022.909570

117. Chen XR, Wang XT, Hao MQ, Zhou YH, Cui WQ, Xing XX, et al. Homology modeling and virtual screening to discover potent inhibitors targeting the imidazole glycerophosphate dehydratase protein in staphylococcus xylosus. *Front Chem.* (2017) 5:98. doi: 10.3389/fchem.2017.00098

118. Yang J, Roy A, Zhang Y. Protein–Ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics.* (2013) 29:2588–95. doi: 10.1093/bioinformatics/btt447

119. Tian W, Chen C, Lei X, Zhao J, Liang J. Castp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* (2018) 46:W363–7. doi: 10.1093/nar/gky473

120. Bhasme PC, Kurjogi MM, Sanakal RD, Kaliwal RB, Kaliwal BB. *In silico* characterization of putative drug targets in *Staphylococcus saprophyticus*, causing bovine mastitis. *Bioinformation.* (2013) 9:339–44. doi: 10.6026/97320630009339

121. Verma S, Pathak RK. Discovery and optimization of lead molecules in drug designing. In: Singh DB, Pathak RK, editors. *Bioinformatics.* Academic Press Elsevier (2022). p. 253–67. doi: 10.1016/B978-0-323-89775-4.00004-3

122. Agnihotry S, Pathak RK, Srivastav A, Shukla PK, Gautam B. Molecular docking and structure-based drug design. In:*Computer-Aided Drug Design.* Singapore: Springer (2020). p. 115–31. doi: 10.1007/978-981-15-6815-2_6

123. Pant S, Verma S, Pathak RK, Singh DB. Structure-Based drug designing. In: Singh DB, Pathak RK, editors. *Bioinformatics.* Academic Press Elsevier (2022). p. 219–31. doi: 10.1016/B978-0-323-89775-4.00027-4

124. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of autodock. *J Mol Recognit.* (1996) 9:1–5. doi: 10.1002/(SICI)1099-1352(199601)9:1<1::AID-JMR241>3.0.CO;2-6

125. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* (2001) 46:3–26. doi: 10.1016/s0169-409x(00)00129-0

126. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *J Med Chem.* (2010) 53:2719–40. doi: 10.1021/jm901137j

127. Baell JB, Nissink JWM. Seven year itch: pan-assay interference compounds (pains) in 2017-utility and limitations. *ACS Chem Biol.* (2018) 13:36–44. doi: 10.1021/acschembio.7b00903

128. Qing X, Lee XY, De Raeymaecker J, Tame JR, Zhang KY, De Maeyer M, et al. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. *J Recept Ligand Channel Res.* (2014) 7:81–92. doi: 10.2147/JRLCR.S46843

129. Tandon H, Chakraborty T, Suhag V. A concise review on the significance of qsar in drug design. *Chem Biomol Eng.* (2019) 4:45–51. doi: 10.11648/j.cbe.20190404.11

130. Tiwari A, Singh S. Computational approaches in drug designing. In: Singh DB, Pathak RK, editors. *Bioinformatics.* Academic Press Elsevier (2022). p. 207–17. doi: 10.1016/B978-0-323-89775-4.00010-9

131. Ganguly B, Rastogi SK, Prasad S. Computational designing of a poly-epitope fecundity vaccine for multiple species of livestock. *Vaccine.* (2013) 32:11–8. doi: 10.1016/j.vaccine.2013.10.086

132. Gebre MS, Brito LA, Tostanoski LH, Edwards DK, Carfi A, Barouch DH. Novel approaches for vaccine development. *Cell.* (2021) 184:1589–603. doi: 10.1016/j.cell.2021.02.030

133. Awasthi A, Sharma G, Agrawal P. Computational approaches for vaccine designing. In: *Bioinformatics*. Elsevier (2022) p. 317–35. doi: 10.1016/B978-0-323-89775-4.00011-0

134. Pathak RK, Lim B, Kim DY, Kim JM. Designing multi-epitope-based vaccine targeting surface immunogenic protein of *Streptococcus agalactiae* using immunoinformatics to control mastitis in dairy cattle. *BMC Vet Res.* (2022) 18:337. doi: 10.1186/s12917-022-03432-z

135. Neethirajan S. The role of sensors, big data and machine learning in modern animal farming. *Sens Bio Sens Res.* (2020) 29:100367. doi: 10.1016/j.sbsr.2020.100367

136. Dumortier L, Guepin F, Delignette-Muller ML, Boulocher C, Grenier T. Deep learning in veterinary medicine, an approach based on CNN to detect pulmonary abnormalities from lateral thoracic radiographs in cats. *Sci Rep.* (2022) 12:11418. doi: 10.1038/s41598-022-149 93-2

137. Kumar I, Singh SP. Machine learning in bioinformatics. In: Singh DB, Pathak RK, editors. *Bioinformatics*. Academic Press Elsevier (2022). p. 443–56. doi: 10.1016/B978-0-323-89775-4.00020-1

138. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol.* (2022) 23:40–55. doi: 10.1038/s41580-021-00407-0

139. Roberts H, Segev A, editors. Animal behavior prediction with long short-term memory. In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA: IEEE (2020). doi: 10.1109/BigData50022.2020.9378184

140. Qiao Y, Clark C, Lomax S, Kong H, Su D, Sukkarieh S. Automated individual cattle identification using video data: a unified deep learning architecture approach. *Front Anim Sci.* (2021) 2:759147. doi: 10.3389/fanim.2021.759147

141. Peng YQ, Kondo N, Fujiura T, Suzuki T, Wulandari, Yoshioka H, et al. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. *Comput Electron Agr.* (2019) 157:247–53. doi: 10.1016/j.compag.2018.12.023

142. Qiao YL, Guo YY, Yu KP, He DJ. C3d-Convlstm based cow behaviour classification using video data for precision livestock farming. *Comput Electron Agr.* (2022) 193:106650. doi: 10.1016/j.compag.2021.106650

143. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv.* (2021) 49:107739. doi: 10.1016/j.biotechadv.2021.107739

144. Benson D, Boguski M, Lipman D, Ostell J. The national center for biotechnology information. *Genomics.* (1990) 6:389–91. doi: 10.1016/0888-7543(90)90583-G

145. UniProt C. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* (2021) 49:D480–9. doi: 10.1093/nar/gkaa1100

146. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res.* (2000) 28:263–6. doi: 10.1093/nar/28.1.263

147. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res.* (2000) 28:235–42. doi: 10.1093/nar/28.1.235

148. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* (2022) 50:D439–44. doi: 10.1093/nar/gkab1061

149. Irwin JJ, Shoichet BK. Zinc–a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* (2005) 45:177–82. doi: 10.1021/ci049714+

150. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* (2002) 30:207–10. doi: 10.1093/nar/30.1.207

151. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.* (2022) 50:D387–90. doi: 10.1093/nar/gkab1053

152. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* (2000) 28:27–30. doi: 10.1093/nar/28.1.27

153. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. String V11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019) 47:D607–13. doi: 10.1093/nar/gky1 131

154. Chelliah V, Laibe C, Le Novere N. Biomodels database: a repository of mathematical models of biological processes. *Methods Mol Biol.* (2013) 1021:189–99. doi: 10.1007/978-1-62703-450-0_10

155. Mishra D, Yadav S, Sikka P, Jerome A, Paul S, Rao A, et al. Snprbb: economically important trait specific snp resources of buffalo (*Bubalus bubalis*). *Conserv Genet Resourc.* (2021) 13:283–9. doi: 10.1007/s12686-021-01210-x

156. Sarika, Arora V, Iquebal MA, Rai A, Kumar D. *In silico* mining of putative microsatellite markers from whole genome sequence of water buffalo (*Bubalus Bubalis*) and development of first buffsatdb. *BMC Genomics.* (2013) 14:43. doi: 10.1186/1471-2164-14-43

157. McCarthy FM, Pendarvis K, Cooksey AM, Gresham CR, Bomhoff M, Davey S, et al. Chickspress: a resource for chicken gene expression. *Database.* (2019) 2019:baz058. doi: 10.1093/database/baz058

158. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. *Nucleic Acids Res.* (2022) 50:D988–95. doi: 10.1093/nar/gkab1049

159. Garcia JF, Carmo AS, Utsunomiya YT, Rezende Neves HH, Carvalheiro R, Tassell CV, et al., editors. How bioinformatics enables livestock applied sciences in the genomic era. In: *Brazilian Symposium on Bioinformatics*. Springer (2012). doi: 10.1007/978-3-642-31927-3_17

160. Bayat A. Science, medicine, and the future-bioinformatics. *BMJ.* (2002) 324:1018–22. doi: 10.1136/bmj.324.7344.1018

161. Council NR. *Critical Needs for Research in Veterinary Science* (2005). Washington, DC: The National Academies Press. Available online at: https://nap.nationalacademies.org/catalog/11366/critical-needs-for-research-in-veterinary-science

162. Leinonen R, Akhtar R, Birney E, Bonfield J, Bower L, Corbett M, et al. Improvements to services at the european nucleotide archive. *Nucleic Acids Res.* (2010) 38:D39–45. doi: 10.1093/nar/gkp998

163. Abouelkhair MA. Non-SARS-CoV-2 genome sequences identified in clinical samples from covid-19 infected patients: evidence for co-infections. *PeerJ.* (2020) 8:e10246. doi: 10.7717/peerj.10246

164. Simons A. *A Quality Control Tool for High Throughput Sequence Data* (2010). Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

165. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

166. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* (2011) 17:10–2. doi: 10.14806/ej.17.1.200

167. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics.* (2018) 34:i884–i90. doi: 10.1093/bioinformatics/bty560

168. Kim D, Langmead B, Salzberg SL. Hisat: a fast spliced aligner with low memory requirements. *Nat Methods.* (2015) 12:357–60. doi: 10.1038/nmeth.3317

169. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. *Bioinformatics.* (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352

170. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* (2012) 9:357–9. doi: 10.1038/nmeth.1923

171. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics.* (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324

172. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from Rna-Seq data without a reference genome. *Nat Biotechnol.* (2011) 29:644–52. doi: 10.1038/nbt.1883

173. Robinson MD, McCarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616

174. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for Rna-Seq data with Deseq2. *Genome Biol.* (2014) 15:550. doi: 10.1186/s13059-014-0550-8

175. Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. *BMC Bioinformatics.* (2008) 9:559. doi: 10.1186/1471-2105-9-559

176. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010) 20:1297–303. doi: 10.1101/gr.107524.110

177. Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. *Genome Res.* (2008) 18:821–9. doi: 10.1101/gr.074492.107

178. Supek F, Bosnjak M, Skunca N, Smuc T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE.* (2011) 6:e21800. doi: 10.1371/journal.pone.0021800

179. Howe EA, Sinha R, Schlauch D, Quackenbush J. Rna-Seq analysis in mev. *Bioinformatics.* (2011) 27:3209–10. doi: 10.1093/bioinformatics/btr490

180. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene Set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* (2005) 102:15545–50. doi: 10.1073/pnas.0506580102

181. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using diamond. *Nat Methods.* (2015) 12:59–60. doi: 10.1038/nmeth.3176

182. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* (2005) 21:3674–6. doi: 10.1093/bioinformatics/bti610

183. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303

184. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. Ucsf chimera–a visualization system for exploratory research and analysis. *J Comput Chem.* (2004) 25:1605–12. doi: 10.1002/jcc.20084

185. Choi Y, Chan AP. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics.* (2015) 31:2745–7. doi: 10.1093/bioinformatics/btv195

186. Csizmadia P. *Marvinsketch and Marvinview: Molecule Applets for the World Wide Web*. (1999). doi: 10.3390/ecsoc-3-01775

187. Daina A, Michielin O, Zoete V. Swissadme: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep.* (2017) 7:42717. doi: 10.1038/srep42717

188. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. Gromacs: fast, flexible, and free. *J Comput Chem.* (2005) 26:1701–18. doi: 10.1002/jcc.20291

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/fvets.2022.1116400

A multiepitope vaccine candidate against infectious bursal disease virus using immunoinformatics-based reverse vaccinology approach

Irfan Gul[1,2], Amreena Hassan[1,2], Jan Mohd Muneeb[1], Towseef Akram[1], Ehtishamul Haq[2], Riaz Ahmad Shah[1], Nazir Ahmad Ganai[1], Syed Mudasir Ahmad[1], Naveed Anjum Chikan[3] and Nadeem Shabir[1]*

[1]Laboratory of Vaccine Biotechnology, Division of Animal Biotechnology, Faculty of Veterinary Sciences and Animal Husbandry, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar, India, [2]Department of Biotechnology, University of Kashmir, Srinagar, India, [3]Division of Computational Biology, Daskdan Innovations Pvt. Ltd., Srinagar, India

Check for updates

OPEN ACCESS

EDITED BY
Rajesh Kumar Pathak,
Chung-Ang University, Republic of Korea

REVIEWED BY
Altaf Hussain,
The Pirbright Institute, United Kingdom
Rajesh Chhabra,
Lala Lajpat Rai University of Veterinary and Animal Sciences, India

*CORRESPONDENCE
Nadeem Shabir
✉ nadeem.shabir@skuastkashmir.ac.in

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

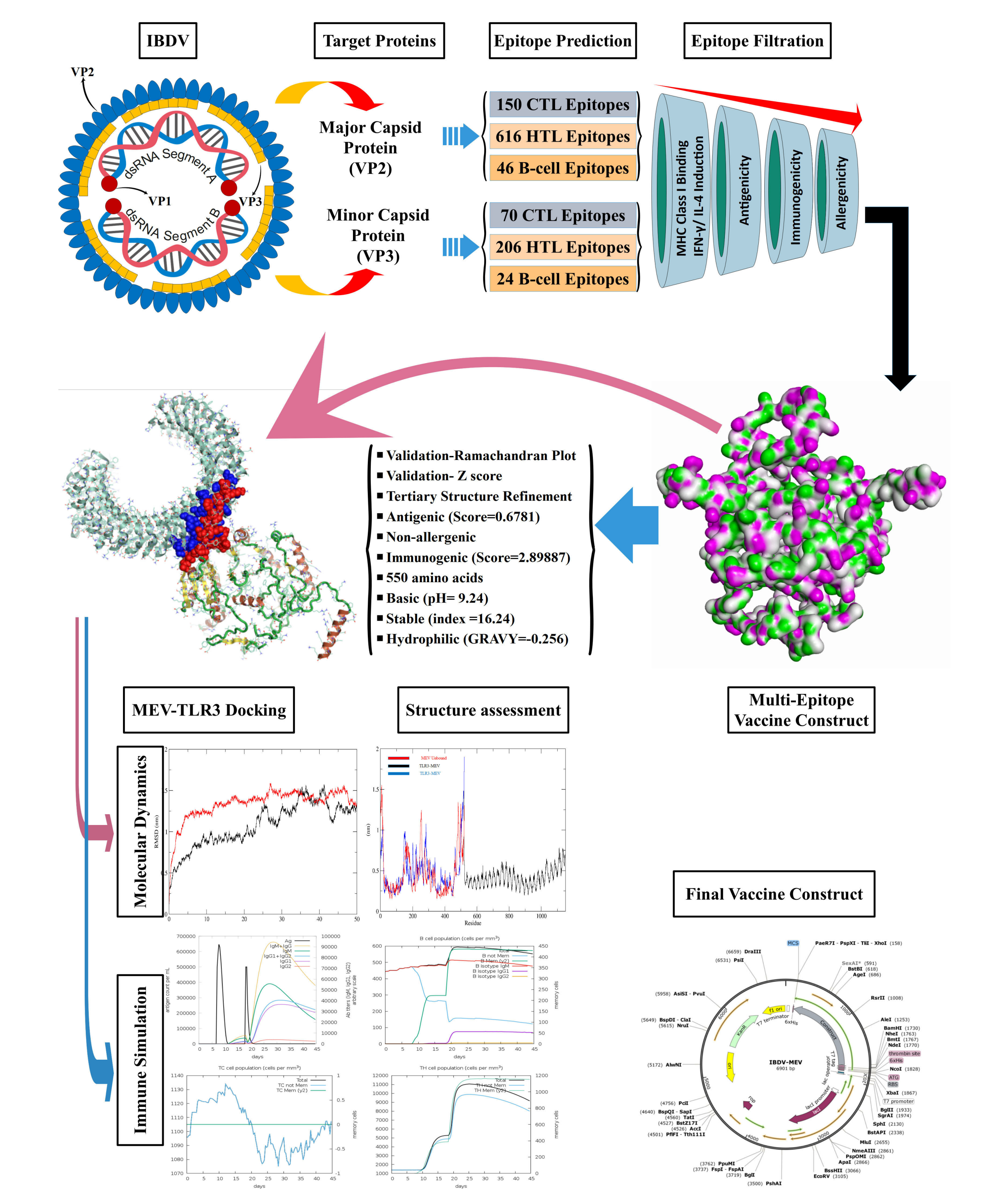
RECEIVED 05 December 2022
ACCEPTED 29 December 2022
PUBLISHED 13 January 2023

CITATION
Gul I, Hassan A, Muneeb JM, Akram T, Haq E, Shah RA, Ganai NA, Ahmad SM, Chikan NA and Shabir N (2023) A multiepitope vaccine candidate against infectious bursal disease virus using immunoinformatics-based reverse vaccinology approach.
Front. Vet. Sci. 9:1116400.
doi: 10.3389/fvets.2022.1116400

Infectious bursal disease virus is the causative agent of infectious bursal disease (Gumboro disease), a highly contagious immunosuppressive disease of chicken with a substantial economic impact on small- and large-scale poultry industries worldwide. Currently, live attenuated vaccines are widely used to control the disease in chickens despite their issues with safety (immunosuppression and bursal atrophy) and efficiency (breaking through the maternally-derived antibody titer). To overcome the drawbacks, the current study has, for the first time, attempted to construct a computational model of a multiepitope based vaccine candidate against infectious bursal disease virus, which has the potential to overcome the safety and protection issues found in the existing live-attenuated vaccines. The current study used a reverse vaccinology based immunoinformatics approach to construct the vaccine candidate using major and minor capsid proteins of the virus, VP2 and VP3, respectively. The vaccine construct was composed of four $CD8^+$ epitopes, seven $CD4^+$ T-cell epitopes, 11 B-cell epitopes and a Cholera Toxin B adjuvant, connected using appropriate flexible peptide linkers. The vaccine construct was evaluated as antigenic with VaxiJen Score of 0.6781, immunogenic with IEDB score of 2.89887 and non-allergenic. The 55.64 kDa construct was further evaluated for its physicochemical characteristics, which revealed that it was stable with an instability index of 16.24, basic with theoretical pI of 9.24, thermostable with aliphatic index of 86.72 and hydrophilic with GRAVY score of −0.256. The docking and molecular dynamics simulation studies of the vaccine construct with Toll-like receptor-3 revealed fair structural interaction (binding affinity of −295.94 kcal/mol) and complex stability. Further, the predicted induction of antibodies and cytokines by the vaccine construct indicated the possible elicitation of the host's immune response against the virus. The work is a significant attempt to develop next-generation vaccines against the infectious bursal disease virus though further experimental studies are required to assess the efficacy and protectivity of the proposed vaccine candidate *in vivo*.

KEYWORDS

IBDV, immunoinformatics, B cell epitopes, T cell epitopes, molecular dynamics simulations



GRAPHICAL ABSTRACT

Introduction

Infectious bursal disease (IBD) is an economically significant and contagious poultry disease. IBD, also known as Gumboro disease, is caused by the double-stranded RNA virus (dsRNA) known as infectious bursal disease virus (IBDV). The virus belongs to the family Birnaviridae, replicates in the bursa of Fabricius (BF) in young chickens causing depletion of B-lymphocytes (1, 2). As a result, young chickens with IBD have significant immunosuppression, putting them at risk to secondary infections (3). IBDV is a nonenveloped virus with icosahedral symmetry and a bi-segmented dsRNA genome (4, 5). Segment A of the genome contains two partially overlapping larger and smaller open reading frames (ORFs). The larger ORF produces a 110 kDa polyprotein that self-processes into two structural capsid proteins (VP2 and VP3) and a non-structural protease protein (VP4), with molecular weights of 48, 33–35, and 24 kDa, respectively (6, 7). The smaller ORF encodes VP5 polypeptide (8), a non-structural protein not required for viral replication *in vitro* but crucial for virus release. The VP2 polypeptide forms the major capsid of IBDV and carries the main immune determinants for eliciting neutralizing antibodies (9). Due to the considerable conservation of the VP2 amino acid sequence across IBDV strains, the linear epitopes have been identified at the residue level. However, the conformation-dependent epitopes are characterized by the core area covering amino acid residues 206–350, the only place where antigenic alterations have been found. The minor capsid protein VP3 is a group-specific immunogenic antigen, with the earliest antibodies appearing after IBDV infection directed at VP3 (10). Segment B of the viral genome encodes for the non-structural protein VP1 (97 kDa), the RNA-dependent RNA polymerase (RdRp) (11). Bound to the genomic RNA, the RdRp stays enclosed within the viral particle.

Adequate control of IBD is possible only by following vaccination regimes as the highly contagious IBDV is a very resilient and persistent virus that survives in poultry houses despite stringent disinfection (12). Despite the many advantages present-day IBD vaccinations (Live attenuated vaccines; LAVs) provide, further improvement is warranted for various reasons. The efficacy of LAVs has been found to decrease in the presence of maternally derived antibodies (MAb) which protect the young chicken during the first few weeks (13, 14). Besides poor efficacy in the presence of MAb, they also possess serious safety issues as they cause varying degrees of bursal atrophy and degeneration as well, in addition to the emergence of antigenic variants in vaccinated flocks, particularly very virulent strains (15–17).

Multiepitope-based vaccines (MEV) are peptide-based vaccines that consist of T cell and B cell epitopes and have the ability to trigger efficient cellular and humoral immune responses (18). MEV can prove a promising strategy for combating viral infections, potentially eliciting a broad immune response due to T cell receptor (TCR) recognized Major Histocompatibility Complex (MHC)-restricted epitopes from target antigens. Moreover, MEV offers improved immunogenicity and long-lasting immune responses without any immunization-related side effects compared to traditional vaccines (19–25). Although the MEV with such advantages have the potential to prove powerful prophylactic and therapeutic agents, the screening of appropriate target antigens and their immunodominant epitopes, as well as the development of an effective delivery system, continue to be the current challenges of MEV design. Therefore, the development of an effective MEV depends on selecting suitable candidate antigens and the immunodominant epitopes associated with them (26–28). Hence this study aimed to develop a potential MEV against IBDV by targeting major and minor capsid proteins through immunoinformatics, molecular modeling and reverse vaccinology approaches.

Materials and methods

The retrieval of protein sequences

The VP2 and VP3 protein sequences from 10 distinct IBDV strains (Supplementary Table 1) were obtained in FASTA format from the National Center for Biotechnology Information (NCBI) protein database (https://www.ncbi.nlm.nih.gov/protein). Multiple sequence alignment was performed on the reference sequences obtained from NCBI using DNA star (DNASTAR, Inc.Madison, WI, USA) with ClustalW parameters. The antigenicity of the reference sequences was evaluated using the VaxiJen v2.0 Server, using a 0.4 antigenicity threshold (http://www.ddg-pharmfac.net/VaxiJen/VaxiJen/VaxiJen.html) (29).

T-cell epitopes Identification

In this study, human HLA alleles were considered instead of chicken HLA alleles because of the unavailability of the applicable data. Consequently, human-related data was utilized to predict the MHC epitopes of selected sequences (30, 31). Humans and chickens have distinct MHC alleles; however, it has been reported that MHC haplotype anchor residue regions in both species are comparable (32).

Cytotoxic T-cell (CTL/CD8$^+$) epitope prediction

The sequences which were found to be antigenic were further submitted to the NetCTL v1.2 server (https://services.healthtech.dtu.dk/service.php?NetCTL-1.2) for the generation of nine amino acid long fragments (33). The fragments were filtered based on interactions with the MHC class I HLA alleles and the production of the CD8$^+$ T cell response. The Stabilized Matrix Base Method (SMM) prediction method of the IEDB tool (http://tools.iedb.org/mhci/) was used to identify the MHC-I HLA binding CTL/CD8$^+$ epitopes out of the resulting nine amino acid long fragments (34). The parameters were set to human as the MHC source species, amino acid length of 9, and IC50 value <250. The screened epitopes were evaluated for antigenicity using the VaxiJen v2.0 server with a 0.5 antigenicity threshold. The potential antigens were further subjected to the IEDB MHC-I immunogenicity tool (http://tools.iedb.org/immunogenicity/) for the evaluation of immunogenicity (35).

Helper T-cell (HTL/CD4$^+$) epitope prediction

The antigenic consensus VP2 and VP3 sequences were submitted to the IEDB MHC-II binding tool (http://tools.iedb.org/mhcii/) to predict HTL epitopes interacting with MHC class II HLA alleles (36). The allele length was adjusted to 15 and the IC50 threshold to 250 to filter out probable epitopes. The screened epitopes were subsequently evaluated for potential IFN-γ cytokine induction using the IFN epitope tool (http://crdd.osdd.net/raghava/ifnepitope/) with SVM (support vector machine) approach and the

IFN vs. non-IFN predictive models (37). Moreover, the IL4 inducer epitopes were identified using the IL4pred tool (http://crdd.osdd.net/raghava/il4pred/) (38). The selected epitopes were then assessed for antigenicity using immunoinformatic techniques identical to those used to test CTL epitopes.

Linear B-cell epitope prediction

The antigenic consensus sequences were subjected to the ABCPRED server (https://webs.iiitd.edu.in/raghava/abcpred/) to identify antigens that can trigger the production of antibodies by eliciting a B cell immune response. The server predicts linear B cell epitopes using an artificial neural network (39). The potential epitopes were screened based on the prediction parameters selected: the window length of 16 and the threshold value of 0.51.

Conservancy and allergenicity assessment

The selected T cell and B cell epitopes determined to be immunogenic and antigenic were evaluated for conservancy using the IEDB conservation across antigen tool (http://tools.iedb.org/conservancy/) (40). The AllerTop v2.0 tool (https://www.ddg-pharmfac.net/AllerTOP/) was used to assess the allergenicity of the conserved epitopes and identify the non-allergic epitopes (41, 42).

Vaccine design and assessment

The top candidates for $CD8^+$, $CD4^+$, and B cell epitopes were identified using several immunoinformatic tools, as indicated above. To construct the IBD-MEV, these epitopes coupled with an adjuvant were linked with appropriate linker peptides. The $CD8^+$/CTL epitopes were linked using an AAY linker, and the $CD4^+$/HTL were connected using GPGPG linkers. The HEYGAEALERAG linker was used to join CTL epitopes with HTL epitopes, while the B cell epitopes were linked by KK linkers. An appropriate adjuvant, cholera toxin subunit B (CTB), was incorporated to the N terminal of the construct peptide using the EAAK linker. The adjuvant was included to enhance the immunogenicity of the vaccine construct (43). The MEV construct was assessed for antigenicity and allergenicity using the VaxiJen v2.0 server and AllerTop v2.0 server, respectively. At the same time, the ProtParam53 web server (https://web.expasy.org/protparam/) determined the physical and chemical characteristics, such as the molecular weight (kDa), the number of amino acid residues, the theoretical isoelectric point (pI), the estimated half-life, the instability index, the aliphatic index, the hydropathicity, and grand average of hydropathicity (GRAVY) (44).

Secondary and Tertiary structure prediction and validation

The primary sequence of the final construct was subjected to the PSIPRED web tool (http://bioinf.cs.ucl.ac.uk/psipred/) for prediction and analysis of the secondary structure (45). While the AlphaFold2-based Colabfold was employed to predict and generate the tertiary structure of the vaccine construct (46, 47). To enhance the quality of the structure, the predicted tertiary structure was subjected to molecular refinement with the aid of the GalaxyRefine server (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE) (48). The resulting models were screened using the GDT-HA, RMSD, and MolProbity scores to choose the most refined model, which was then verified using the Ramachandran plot and ProSA-web-predicted Z-score (https://prosa.services.came.sbg.ac.at/prosa.php) (49, 50).

Docking and molecular dynamic simulation analysis

The vaccine construct was docked with the Toll Like Receptor 3 (TLR3; PDB ID: 1ZIW) using the HDOCK server (http://hdock.phys.hust.edu.cn/) with default parameters (51). The server provides 10 poses for each docking run, wherein the model with the lowest binding energy was selected and visualized by PyMOL (https://pymol.org/) and Discovery Studio Biovia 2021 (https://discover.3ds.com/). Molecular Dynamics Simulation by GROMACS 2021.1 was performed using OPLS-AA/L all-atom force field to study the stability of the complex (52). The complex was placed in a unit cell, defined as a 1-nm cube, solvated with water using a solvate model. Ions were added according to the charge present on the vaccine construct, and the obtained electro-neutral structure was relaxed through energy minimization. The equilibrating of the water around the complex was conducted under NVT and NPT conditions for 100 ps. The temperature was set to a maximum of 300 K. Following the equilibration phases, MD simulation data was collected to perform the 50 ns final run with a time step of 2 fs at constant pressure (1 bar) and temperature (300 K). The resulting trajectories were analyzed using the inbuilt utilities of GROMACS.

In silico cloning and optimization of vaccine construct

Using the Java Codon Adaptation Tool (http://www.jcat.de/), the vaccine construct was codon optimized using the *Escherichia coli* K12 strain as the host organism. The JCat adaptation was defined using the codon adaptation index (CAI) and GC content of the optimizes sequence (53). The ideal CAI score of an edited gene sequence is around 0.8 and 1.0, with a GC percentage of 30%–70%, suggesting better gene expression in the associated organism with no translation mistakes (54). To the optimized vaccine sequence, restriction sites *BamHI* (GGATCC) was added to the $5^{'}$ end while *XhoI* (CTCGAG) restriction site was added to the $3'$ end. SnapGene Viewer V3.2.1 (http://www.snapgene.com/) carried out the *in-silico* cloning of the optimized vaccine sequence into the pET-28a (+) expression vector system.

In silico immune simulation of vaccine construct

The C-ImmSim server (https://kraken.iac.rm.cnr.it/C-IMMSIM/) was used to model and evaluate the immune response of the vaccine construct for the specified vaccination program (55). A vaccination without lipopolysaccharide (LPS) was used for the simulation, and all other parameters were left at their default values. A single injection of the vaccine construct was administered at two intervals; Day 7 and Day 18.

Results

Cytotoxic T-cell (CTL/CD8$^+$) epitope prediction

The VP2 and VP3 consensus sequences were subjected to the NetCTL v1.2 server to predict specific immunogenic CTL epitopes. A total of 150 and 70 nonamer epitopes were obtained from VP2 and VP3 proteins, and each of them had a considerable binding affinity for the 12 superfamily HLA alleles. Using the IEDB MHC-I prediction tool, the CTL epitope nonamers were scrutinized for specific MHC-I binding affinity with the SSM-based method. The epitopes were filtered by the IC50 value parameter (<250), yielding 86 VP2 and 37 VP3 CTL epitopes. The screened epitopes were examined with the VaxiJen v2.0 server for antigenicity (threshold ≥ 0.5). Among the predicted epitopes of VP2 and VP3 proteins, 47 and 16 epitopes showed considerable antigenic potential, with the highest antigenic score of 1.6076 and 0.8179 for VP2 epitope "TSYDLGYVR" and VP3 epitope "EAAANVDPL." Following the assessment of immunogenicity using the IEDB tool, the epitopes were filtered to 26 VP2 and 8 VP3 immunogenic epitopes. Finally, the allergenicity and conservancy analysis was carried out using the AllerTop v2.0 server and IEDB conservation across antigen tool, where the allergenic and non-conserved epitopes were screened out, and only 15 VP2 and 2 VP3 CTL epitopes were regarded as concluding predicted epitopes (Table 1).

Helper T-cell (HTL/CD4$^+$) epitope prediction

Overall, 616 VP2 and 206 VP3 15-mer HTL epitopes were identified using IEDB MHC-II binding tool screening out *via* filtration based on IEDB tool IC50 value (<250) and VaxiJen tool antigenicity score (≥ 0.4). The 15-mer epitopes were further examined for IFN-gamma and interleukin inducer properties using IFN epitope and IL-4pred immunoinformatic tools. A total of 99 VP2 and 100 VP3 CD4$^+$ T cell epitopes exhibited the property to induce IFN-γ, while only 25 VP2 epitopes and 16 VP3 epitopes exhibited IL-4 inducer properties. Finally, the antigenicity and allergenicity analysis was carried out using VaxiJen and AllerTop v2.0 servers, where the non-antigen and allergenic epitopes were screened out. 6 VP2 epitopes and 1 VP3 epitope were concluded as the most promising HTL epitope candidates for the final vaccine construct (Table 2).

Linear B-cell epitope prediction

An iitd.edu.in server was used to generate 46 VP2 and 24 VP3 B cell epitopes. Out of these, 28 VP2 and 10 VP3 epitopes were revealed as antigenic by the VaxiJen server (threshold >0.5). The Immunogenicity analysis further filtered the epitopes to 14 VP2 and 4 VP3 Immunogenic epitopes. Among the immunogenic epitopes, only 9 VP2 and 3 VP3 B cell epitopes were assessed as non-allergenic and selected for the final vaccine construct, omitting the allergenic epitopes (Table 3).

Vaccine design and assessment

The epitopes were combined to construct the MEV candidate against IBDV based on their antigenicity, immunogenicity, non-allergenic and non-overlapping characteristics. The final IBD-MEV design included 4 CTL, 7 HTL, 11 linear B cell epitopes, and a CTB adjuvant, with AAY linkers connecting the CTL, GPGPG linkers connecting the HTL, and KK linkers connecting the B cell epitopes. The CTB adjuvant was attached in the N-terminal by an EAAK linker to increase the immunogenicity of IBD-MEV. Moreover, the HEYGAEALERAG linker was inserted between CTL and HTL epitopes, and an EAAK liner was added to the C-terminal of the IBD-MEV construct (Figure 1A). The 522-residue IBD-MEV construct with a molecular weight of 55.64 kDa was evaluated for antigenicity, immunogenicity, and allergenicity, in addition to physical and chemical properties. The vaccine was demonstrated as antigenic (VaxiJen score = 0.6781), immunogenic (score = 2.89887), and non-allergenic. The assessment of the physicochemical properties using the ProtParam server presented that the IBD-MEV construct has a theoretical isoelectric point (PI) of 9.24, making it substantially basic. The IBD-MEV construct was determined to be stable with an instability index of 16.24, thermostable with an aliphatic index of 86.72, and hydrophilic with GRAVY scores of −0.256.

Secondary and tertiary structure prediction and validation

The PSIPRED server examined the secondary structural properties of the IBD-MEV. Accordingly, the construct had 22.22% of the amino acids in the α-helix conformation and 23.56% of amino acids in the β-strand conformation and 54.22% in coil structure conformations (Supplementary Figure 1). The tertiary structure of the IBD-MEV construct was predicted using the AlphaFold2-based Colabfold, while the GalaxyRefine server was employed to refine the structure (Figure 1B). The refined model was obtained with a GDT-HA score of 0.8582, an RMSD value of 0.682, a MolProbity score of 1.459 and a rotamer score of 0.7, indicating the high quality of the model. The preferred refined model structure was validated using the Ramachandran plot, with 96.0% residues in the favored region (Figure 1C). The model was further validated using ProSA-web and has a half-life of 30 h in mammalian reticulocytes (*in vitro*), >20 h in yeast, and >10 h in *E. coli* (*in vivo*). Moreover, a *Z*-score of −4.49 was obtained, signifying the high quality of the structure (Figure 1D). The structural assessment of the IBD-MEV tertiary structure is displayed in Supplementary Figures 1B–F).

Docking and molecular dynamic simulation analysis

The docking of the IBD-MEV construct was performed with TLR3 as a receptor using the HDOCK server. TLR3 is a significant

TABLE 1 Final predicted cytotoxic T cells (CD8$^+$/CTL) epitopes.

	Epitopes	Position	HLA allele	ic50	Immunogenicity	Antigenicity	Allergenicity
VP2	KTVWPTREY	18	HLA-A*30:02	117.8853942	0.36421	0.6857	Non-allergen
			HLA-B*15:02	123.6602158			
	LKIAGAFGF	124	HLA-B*15:02	191.9685126	0.271	0.9619	Non-allergen
	VLVGEGVTV	29	HLA-A*02:01	197.2922383	0.23442	0.5956	Non-allergen
			HLA-A*02:06	66.76517628			
	GIKTVWPTR	47	HLA-A*31:01	51.8167297	0.23109	0.8973	Non-allergen
	YGRFDPGAM	90	HLA-B*15:02	114.3483757	0.19722	1.1414	Non-allergen
			HLA-B*35:01	147.4280088			
	RLGDPIPAI	23	HLA-A*02:01	113.7915466	0.1613	0.7382	Non-allergen
			HLA-A*02:06	194.3345415			
	SYDLGYVRL	54	HLA-B*15:02	47.44932078	0.09064	1.5198	Non-allergen
	TSYDLGYVR	49	HLA-A*31:01	16.27308901	0.06322	1.6076	Non-allergen
			HLA-A*68:01	16.92349276			
			HLA-A*11:01	87.96299338			
	GEGVTVLSL	108	HLA-B*40:02	133.0025599	0.02318	0.5789	Non-allergen
			HLA-B*15:02	179.982384			
			HLA-B*40:01	58.1969199			
VP3	KVYEVNHGR	16	HLA-A*68:01	54.88959167	0.18076	0.773	Non-allergen
			HLA-A*11:01	110.9941271			
			HLA-A*31:01	7.878980035			
	EAAANVDPL	29	HLA-A*68:02	35.52957397	0.09687	0.8179	Non-allergen
			HLA-B*15:02	54.47910949			
			HLA-B*35:01	125.1929866			

The epitopes were predicted using NetCTL v1.2 and scrutinized using the IEDB MHC-I prediction tool.

TABLE 2 Final predicted Helper T cells (CD4$^+$/HTL) epitopes.

	Epitopes	HLA allele	ic50	Immunogenicity	Antigenicity	Allergenicity
VP2	SEITQPITSIKLEIV	HLA-DRB1*07:01	112	0.03638	0.5556	Non-allergen
		HLA-DRB1*01:01	183			
	LGYVRLGDPIPAIGL	HLA-DRB1*01:01	151	0.41149	1.1488	Non-allergen
	DLGYVRLGDPIPAIG	HLA-DRB1*01:01	170	0.36082	1.3044	Non-allergen
	YDLGYVRLGDPIPAI	HLA-DRB1*01:01	173	0.28928	1.3085	Non-allergen
	TSYDLGYVRLGDPIP	HLA-DRB1*01:01	181	0.2418	1.4101	Non-allergen
	SYDLGYVRLGDPIPA	HLA-DRB1*01:01	186	0.25524	1.3072	Non-allergen
VP3	ELESAVRAMEAAANV	HLA-DRB1*04:04	37	0.0635	0.4095	Non-allergen
		HLA-DRB1*04:01	172			
		HLA-DRB1*01:01	37			

The epitopes were predicted using IEDB MHC-II prediction tool and screened as IFN-gamma and interleukin inducers.

TLR family member recognizing viral double-stranded RNA. The produced docked models were visualized using the PyMOL and Discovery Studio Biovia 2021. The HDOCK returned models were screened based on the binding affinity, and the model with ΔG value of -295.94 kcal/mol was selected (Figure 2). The interacting residues of TLR3 and MEV reveal various types of interaction between the two structures (Figure 2C). The complex was subjected to MD simulations to assess the docked complex's stability, binding and dynamics (Supplementary Movie). The backbone RMSD and residue-wise RMSF trajectories were analyzed throughout the 50

TABLE 3 Final selected linear B-cell epitopes.

	Epitopes	Position	Immunogenicity	Antigenicity	Allergenicity
VP2	DRLGIKTVWPTREYTD	398	0.43501	0.736	Non-allergen
	GYVRLGDPIPAIGLDP	168	0.42048	0.8952	Non-allergen
	NLTVGDTGSGLIVFFP	38	0.3763	0.8846	Non-allergen
	GSVVTVAGVSNFELIP	352	0.35349	0.9296	Non-allergen
	KNLVTEYGRFDPGAMN	373	0.34016	0.5987	Non-allergen
	LILSERDRLGIKTVWP	392	0.19243	1.0791	Non-allergen
	GLTAGTDNLMPFNIVI	273	0.18982	0.8909	Non-allergen
	NSPLKIAGAFGFKDII	425	0.16852	0.7304	Non-allergen
	TSEITQPITSIKLEIV	290	0.15834	0.5207	Non-allergen
VP3	GVEARGPTPEGAQREK	100	0.33321	0.5198	Non-allergen
	TPEWVALNGHRGPSPG	132	0.30795	0.7318	non-allergen
	PTPEGAQREKDTRISK	106	0.11713	0.5853	Non-allergen

The epitopes were predicted using the abcpred web tool.

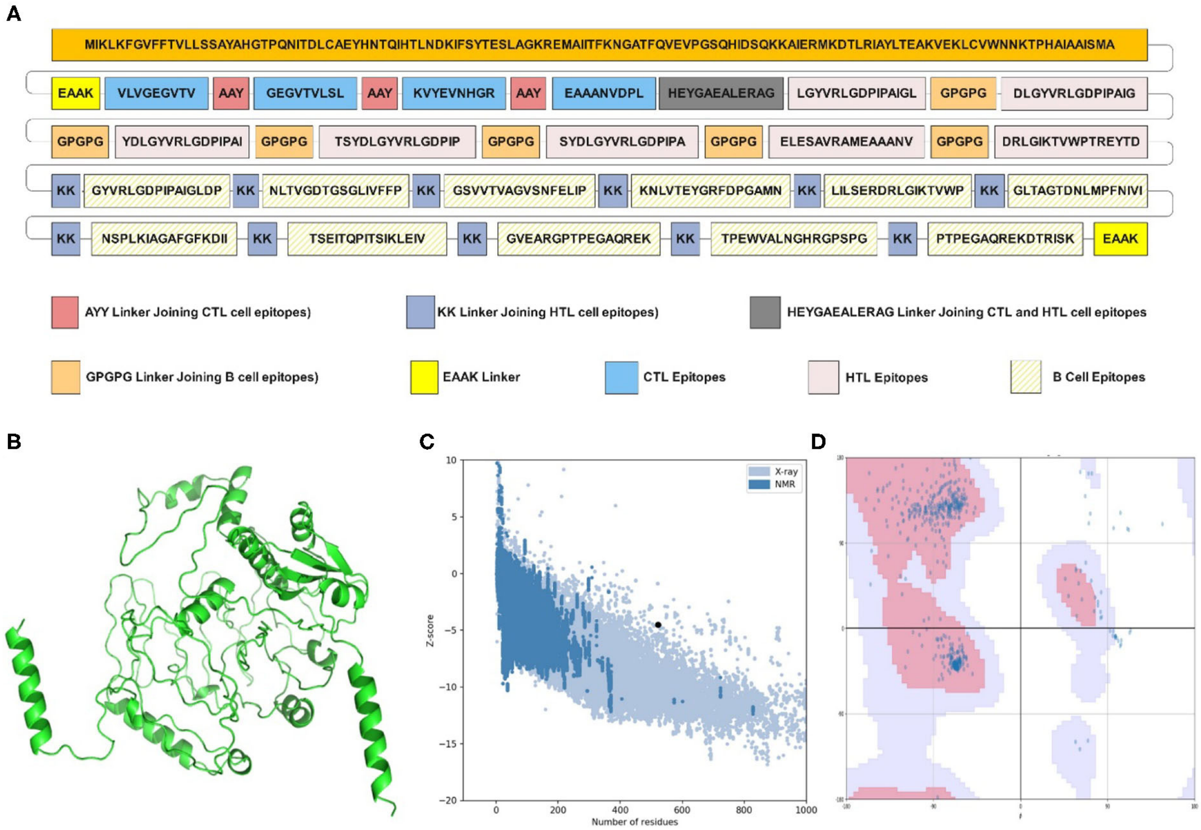


FIGURE 1
Structural analysis and validation of designed vaccine. **(A)** Schematic design of the final vaccine construct. AAY Linkers join the CTL epitopes, HTL epitopes are joined by GPGPG linkers and B-cell epitopes by KK linkers. The Cholera Toxin B (CTB) adjuvant is added to the N-terminus of the sequence by an EAAK linker. An additional EAAK linker C-Terminal and HEYGAEALERAG linkers between CTL and HTL epitopes were incorporated. **(B)** The refined three-dimensional structure of vaccine construct; **(C)** ProSA-web assessment of the vaccine tertiary structure. The evaluation revealed a *Z*-score of −4.49, indicating good quality. **(D)** Ramachandran plot analysis of the refined structure. The evaluation revealed that 96.0% of the residues of the vaccine are present in the favored region.

ns simulation. The comparison of RMSD fluctuation for backbone atoms of IBD-MEV before docking and MEV-TLR3 complex signifies the stability of the MEV system due to the binding of MEV to the TLR3 (Figure 3A). RMSF analysis (Figure 3B) revealed slight fluctuation in docked complex side-chain atoms, which may reflect high interaction between the IBD-MEV and TLR3.

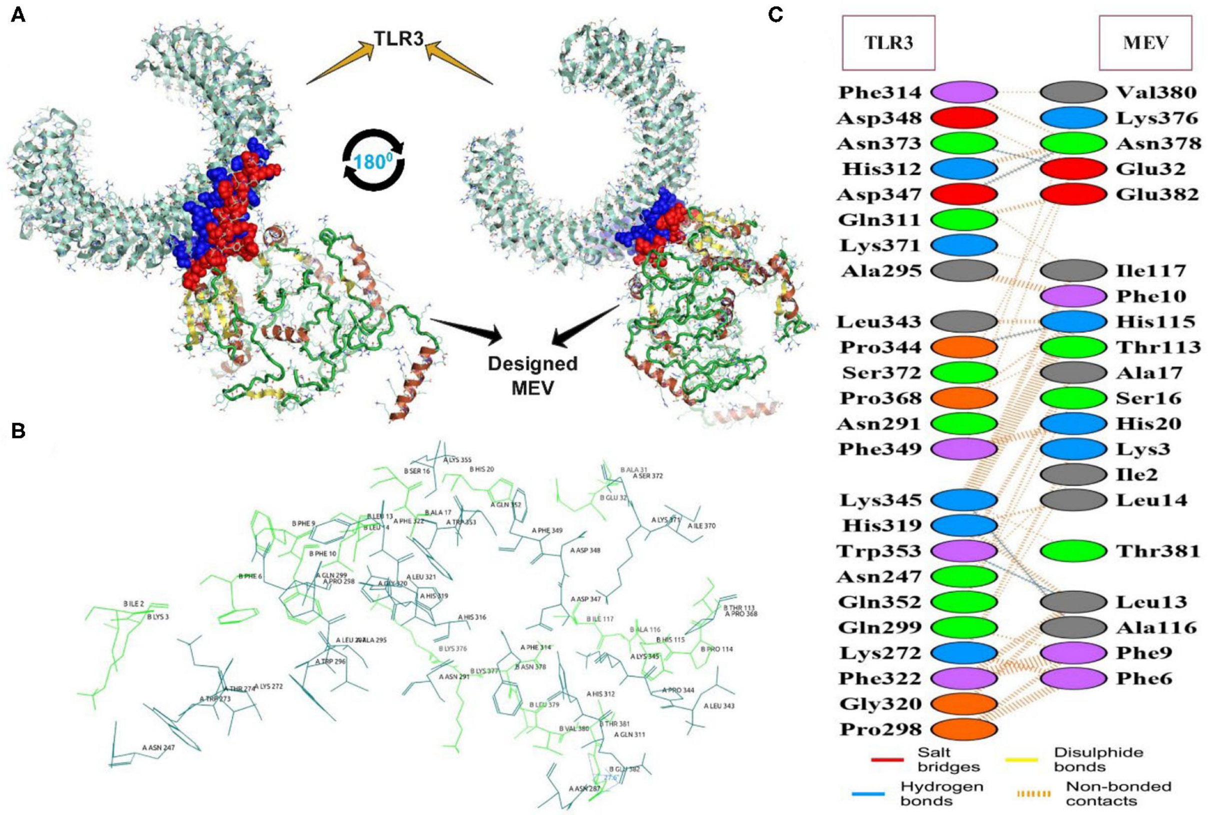


FIGURE 2
The molecular interaction analysis of the designed MEV with TLR3 after protein-protein docking. **(A)** Interacting tertiary structure whereby interacting residues are shown by blue (TLR3) and red (MEV); **(B)** The interacting residues; TLR3 (Teal) and MEV (Green); **(C)** Different interaction between the interacting residues of TLR3 and MEV.

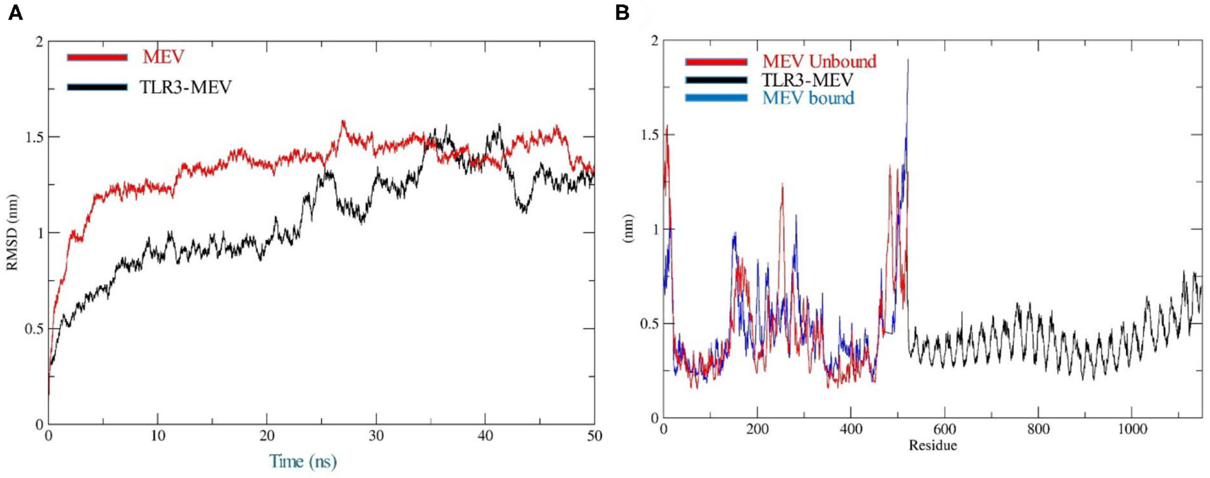


FIGURE 3
Molecular dynamics simulation analysis at 50-ns MD simulation. **(A)** Analysis of RMSD trajectories for MEV-TLR3 complex (Black) and MEV (Red), relative to the backbone. The RMSD plot showed structural stability of the complex with minimum deviations; **(B)** Analysis of RMSF trajectories for MEV-TLR3 complex (Black), MEV (Red) and MEV bound (Blue) to TLR3. The RMSF plot shows the flexibility of interacting side-chain regions.

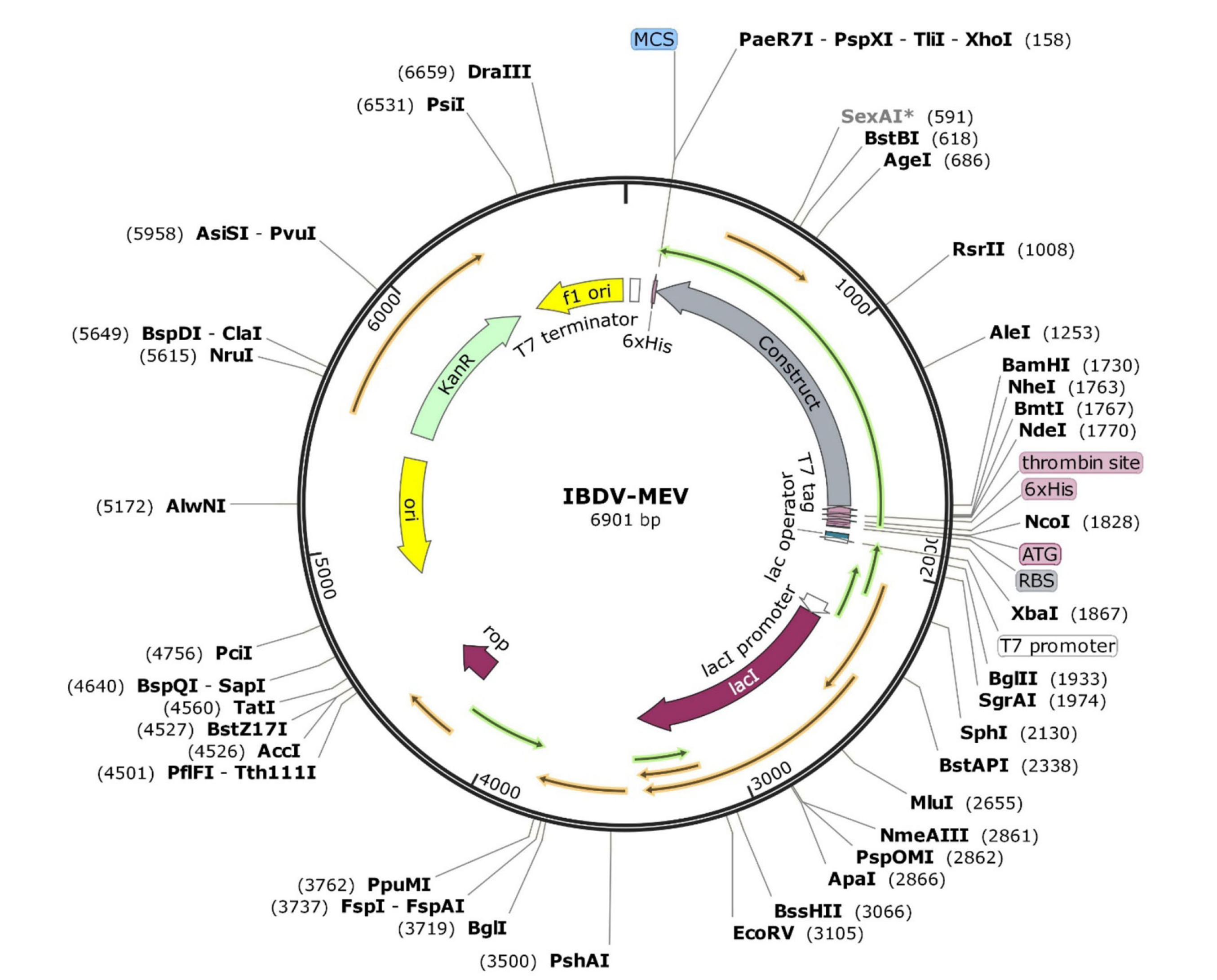


FIGURE 4
In silico restriction cloning of the optimized sequence of the designed MEV into the pET28a(+) expression vector. The MEV construct is labeled as *Construct,* and the restriction sites are incorporated at N-terminal (*XhoI*) and C-terminal (*BamHI*) of the vaccine construct.

In silico cloning and optimization of vaccine construct

The JCat server optimized the codon usage for maximal expression of the IBD-MEV construct according to *E. coli* (strain K12). The obtained CAI value of 0.99 and GC-content of 52.36% imply the effectiveness of IBD-MEV expression in the selected host. The predicted DNA sequence of the IBD-MEV construct was cloned into the pET-28a(+) expression system using the SnapGene. *BamHI* restriction sequence was incorporated at the N-terminal, and *XhoI* site was incorporated at the C-terminal of the construct (Figure 4).

In silico immune simulations of vaccine construct

Immunological simulation findings confirmed various immune profiles created by the vaccination, with the vaccine inducing an immune response *via* an increase in antibodies after delivery to the simulation. The vaccine doses were administered in two intervals: the first dose for a 7-day old chick and the second after 11 days of the first dose (Day 18). The immune response was studied for 45 days. With C-ImmSim simulation, in comparison to the primary reaction indicated by IgM, delivery of the IBD-MEV construct resulted in a considerable increase in the tertiary immune response. After receiving the vaccination, the B cell population produced memory cells that would keep the memory if the host became reinfected (Figure 5). The existence of antibodies that successfully preserved the likelihood of an antigenic rush was confirmed by the drop in antigen level with each vaccination.

Discussion

IBDV is one of the top infectious issues affecting young chickens, with a significant socio-economic impact on the poultry industry with direct and indirect losses (5). Direct losses include morbidity and

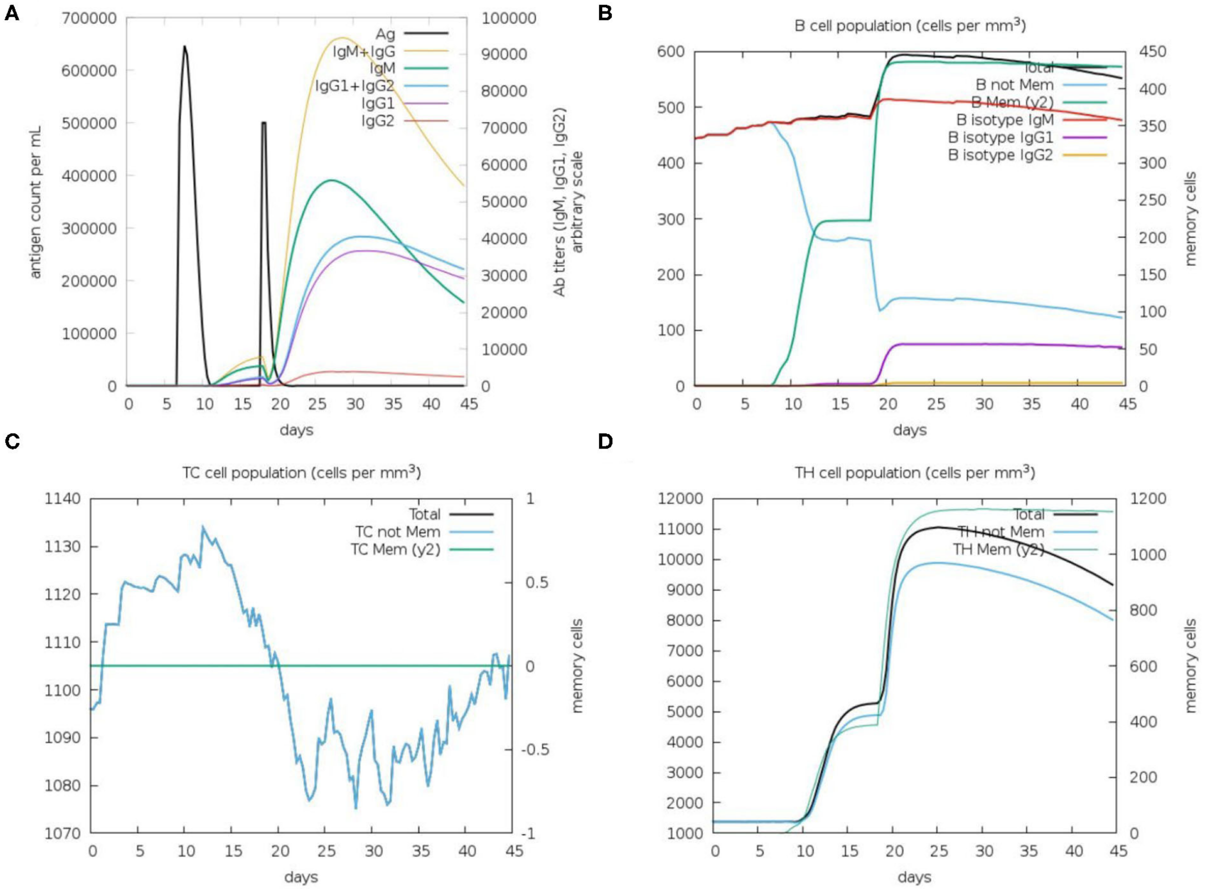


FIGURE 5
C-ImmSim *in silico* immune simulation analysis, showing immune response against IBDV-MEV construct. **(A)** Immunoglobin production (colored peaks) in response to vaccine injections (black; 7 and 18 Day); **(B)** Amount of B lymphocytes composed of B memory (y2) and B-isotypes (IgM, IgG1, and IgG2) **(C)** CD8+ T-cytotoxic lymphocytes (CTL) cell populations and **(D)** CD4+ T-helper lymphocytes (HTL) cell population; per state in response to antigen injection.

mortality losses, while indirect losses owe to immunosuppression-induced secondary infections. Since the virus targets the B cells in the BF, chickens typically display immunosuppression, are less responsive to vaccination campaigns, and are more vulnerable to secondary infections. The use of a live attenuated (mild strain) of IBDV is a frequent IBDV vaccination regimen. LAVs imitate infection to induce host immunity and reduce clinical illness or immunosuppression. Even though this treatment prevents clinical indications of the disease, it produces bursal damage. Moreover, there is a risk that the LAVs may revert to a virulent strain, resulting in bursal injury and immunosuppression (15, 16). LAVs are also ineffective against vvIBDV and rapidly neutralized by MAb (30). Recently, the focus has shifted to developing epitope-based vaccines due to their superior safety profiles and logistical manageability. The potential benefits of epitope-based vaccination are improved safety, time-saving, ability to focus on conserved epitopes and specifically engineer epitope combinations for increased potency (56). As a result, rational selections are made to isolate and separate the ingredients needed for the intended immune response using immuno-informatics approaches to vaccine development. The immunoinformatics techniques can be employed to design proper protein antigens that elicit antibody response and cell-mediated immunity. Therefore, this study aimed to construct a potential MEV against IBDV by focusing on the two capsid proteins of the virus, VP2 and VP3. The trimeric form of VP2 makes up the IBDV virion's major capsid, while the dimeric VP3 subunits make up the inner minor capsid. VP2 is preferably targeted in IBDV vaccine development strategies because it is essential for selection, entry into target cells, and induction of protective, neutralizing antibodies (9). VP3 has also been identified as a putative antigen for the production of a multiepitope vaccine (10). The MEVs would mitigate any potential negative consequences of employing the entire virion, reducing the likelihood of reversion to virulence and other vaccine-related adverse effects.

T cell epitopes are antigenic peptides recognized by the TCR when bound to MHC molecules. MHC class I presents CTL, and MHC class II presents HTL epitopes recognized by $CD8^+$ and $CD4^+$ T cells, respectively. After identifying the target epitope, $CD8^+$ T cells mature into CTL, which can destroy malignant or infected cells. In contrast, $CD4^+$ T cells mature into HTL, stimulating B cells to generate antibodies and macrophages to eliminate the target antigen. T lymphocytes are essential immune system cells reportedly required for complete protection and generation of protective antibodies against virulent IBDV (57). The B cell antigenic epitopes are identified by secretory antibodies or B cell receptors to stimulate an immune response (57). Therefore, B cell epitopes are essential to induce humoral or antibody-mediated immunity, which serves as the main line of defense against severe IBDV. This approach used standard servers to identify and evaluate appropriate CTL, HTL and B cell epitopes from the targeted VP2 and VP3 proteins. Based on the evaluations, four CTL, seven HTL and 11 B cell epitopes were selected for the final vaccine construct.

The present investigation added CTB mucosal adjuvant to the final IBD-MEV construct. Adjuvants have become an essential component of most vaccines, enhancing the cell-mediated immune responses, decreasing the antigen dosage, inducing prolonged immune responses and acting as agonists for TLRs (58). The non-toxic CTB has a strong affinity for the gut mucosal GM1-ganglioside receptor (43). CTB has been used extensively in mucosal immunization strategies as a DNA vaccine adjuvant. These strategies have shown CTB as an effective adjuvant for developing mucosal antibody responses and specific immunity. Additionally, CTB activated the signaling pathways through TLR's, which are crucial in connecting innate and adaptive immunity.

To complete the final stage of the IBD-MEV construction, the epitopes and CTB adjuvant were linked using suitable flexible linkers. Linkers are crucial for improving the stability and expression of proteins in developing MEVs. The AAY linkers joining CTL epitopes enhance dissociation and epitope identification by preventing the formation of junctional epitopes. The glycine-rich GPGPG linkers that connect the HTL epitopes enhance the construct's solubility, accessibility, and flexibility of adjacent domains (30). The CTL epitopes were paired with the HTL epitopes using HEYGAEALERAG linkers, which enhance epitope presentation by creating distinct proteasomal and lysosomal cleavage sites. The bi-lysine linker that joined the B cell epitopes helps in the specific presentation of each peptide to antibodies and preserves their individual immunogenic properties (30). A rigid EAAK linker forming an alpha helix connected the CTB adjuvant to the N-terminus of the constructs to improve domain independence and stability (59). The overlapping epitope sequences were scrutinized and merged into one.

In order to confirm that the IBD-MEV construct provides an efficient immune response without eliciting allergic reactions, it is imperative to evaluate the antigenic, immunogenic and allergenic properties. The IBD-MEV was determined to be immunogenic, antigenic and non-allergenic. Generally, a promising vaccine candidate should have a molecular weight lesser than 110 kDa and an instability index lesser than 40, which classify them as relatively stable. The IBD-MEV had a molecular weight of 55.64 kDa and an instability index of 16.24, which meets the criteria for a stable vaccine. While the predicted theoretical pI of 9.24 indicates the basic nature. This may be because IBD-MEV contains basic amino acids such as arginine (4.2%), histidine (1.5%) and lysine (8.6%). The aliphatic index, which is the proportional volume occupied by the protein's aliphatic side chains, determines the thermal stability of a vaccine construct, with higher aliphatic index values indicating thermostability over a wide temperature range (60, 61). The projected aliphatic index of 86.72 for the constructed multiepitope vaccine suggested the thermostability of the protein. A key factor used to assess the protein's solubility is the Grand Average of Hydrophobicity Index (GRAVY), which represents the hydrophobicity value of a peptide. When the GRAVY value is positive, it shows hydrophobicity; when it is negative, it suggests hydrophilicity (61). The construct, with a GRAVY index of −0.256, reflects the polarity and high solubility of the IBD-MEV construct.

The IBD-MEV tertiary structure was modeled using Colabfold, a rapid protein structure and complexes prediction tool based on AlphaFold2 artificial intelligence (AI) system (46, 47). In order to predict a structure close to the native system, the 3D model needs to be refined and validated, which was achieved through the GalaxyRefine server in the study (48). The good quality of the refined model is indicated by the model's global distance test-high accuracy (GDT-HA) score, RMSD value, MolProbity score, and rotamers score. The refined model evaluated using the Ramachandran diagram showed that most of the vaccine's amino acids (96% residues) were located in the favored region. While the ProSA online server's evaluation of a *Z* score supported the vaccine's overall quality (49).

TLRs are conserved membrane-spanning proteins that function as the body's first line of defense and are essential to the innate immune system. TLRs control the transcriptional expression of cytokines by identifying pathogen-associated molecular patterns derived from pathogens (62). The cytokines production triggers the host's innate immune system to mediate antimicrobial response. Among the chicken TLRs, TLR3 tends to recognize viral dsRNA; therefore, IBD-MEV was docked against TLR3 using the HDOCK server (51, 63). The server predicted a robust interaction with a negative Gibbs-free (ΔG) value. The Gibbs free energy is essential to characterize the magnitude of an interaction occurring under certain circumstances in a cell. The more negative the value of the Gibbs free energy, the more energetically feasible the interaction is. Accordingly, a ΔG value of −295.94 kcal/mol indicates stable binding of IBD-MEV and TLR3. PDBSum revealed the existence of H-bond and salt bridge interactions between the IBD-MEV and TLR3 (64, 65). Additional validation of the docking results was performed using 50 ns molecular dynamics simulation analysis, where the root mean square deviation (RMSD) and root mean square fluctuation (RMSF) of the complex, bound and unbound IBD-MEV were determined. RMSD calculates the degree of deviation for a group of atoms to the respective initial reference structure. Thus, high RMSD values would be associated with instability in the structure. The complex structure exhibited lower RMSD trajectories as compared to MEV, indicating that IBD-MEV and TLR3 were bound in a stable and confined manner. RMSF provides more insights regarding the stability of the complex. The bound and unbound MEV structure displayed fluctuations in RMSF analysis, which may be intrinsic to the structure. These findings imply that the IBD-MEV can efficiently activate TLR3 and enhance immune defenses against the IBDV.

By modeling the host's immunological response following vaccination, immune simulations give insight into the capability of the vaccine construct against the pathogen (66). The *in silico* immune simulation results demonstrated the production of memory B cells, T cells and elevated Immunoglobulin (Ig's) levels. Upon the first IBD-MEV administration, modest production of antibodies was simulated, whereas elevated production of antibodies was observed upon the second dose. Among the immunoglobulins, high IgG and IgM levels were simulated, constituting the primary response against the virus. In addition, IgG1 + IgG2, IgG1 and IgG2 comprising the secondary and tertiary response were also noted on vaccine administration. The antigen exposure increased the B lymphocyte count, particularly the memory B lymphocytes. The progressive rise in memory B lymphocytes and immunoglobulins with repeated administration of the antigen confirms the efficacy of IBD-MEV when the host is exposed over a prolonged period of time. The T helper (TH) response exhibited a similar response, with antigen exposure increase in memory cell count was predicted. In contrast, cytotoxic T cells maintained a modest level throughout the simulation. In this way, the IBD-MEV administration simulated an efficient humoral and cell-mediated immune response against IBDV. However, further research and experimental validation of the current study's findings are necessary to confirm and validate the safety, protectivity and efficacy parameters.

Data availability statement

Publicly available datasets were analyzed in this study. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

IG and NS contributed to the conception and design of the study. IG and AH wrote the first draft of the manuscript. JM and TA wrote sections of the manuscript. NC supervised the *in-silico* experiments. EH, RS, NG, and SA contributed to the manuscript reading and revision. All authors approved the submitted version.

Conflict of interest

NC was employed by Daskdan Innovations, PVT Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.1116400/full#supplementary-material

References

1. McFerran JB, McNulty MS, McKillop ER, Connor TJ, McCracken RM, Collins DS, et al. Isolation and serological studies with infectious bursal disease viruses from fowl, turkeys and ducks: demonstration of a second serotype. *Avian Pathol.* (1980) 9:395–404. doi: 10.1080/03079458008418423

2. Jackwood DJ, Saif YM, Hughes JH. Characteristics and serologic studies of two serotypes of infectious bursal disease virus in turkeys. *Avian Dis.* (1982) 26:871–82. doi: 10.2307/1589875

3. Käufer I, Weiss E. Significance of bursa of Fabricius as target organ in infectious bursal disease of chickens. *Infect Immun.* (1980) 27:364–7. doi: 10.1128/iai.27.2.364-367.1980

4. Dobos P, Hill BJ, Hallett R, Kells DT, Becht H, Teninges D, et al. Biophysical and biochemical characterization of five animal viruses with bisegmented double-stranded RNA genomes. *J Virol.* (1979) 32:593–605. doi: 10.1128/jvi.32.2.593-605.1979

5. Kibenge FS, Dhillon AS, Russell RG. Biochemistry and immunology of infectious bursal disease virus. *J Gen Virol.* (1988) 69(Pt 8):1757–75. doi: 10.1099/0022-1317-69-8-1757

6. Böttcher B, Kiselev NA. Stel'Mashchuk VY, Perevozchikova NA, Borisov AV, Crowther RA. Three-dimensional structure of infectious bursal disease virus determined by electron cryomicroscopy. *J Virol.* (1997) 71:325–30. doi: 10.1128/jvi.71.1.325-330.1997

7. van den Berg TP, Eterradossi N, Toquin D, Meulemans G. Infectious bursal disease (Gumboro disease). *Rev Sci Tech.* (2000) 19:509–43. doi: 10.20506/rst.19.2.1227

8. Mundt, E., Beyer, J., and Müller, H. Identification of a novel viral protein in infectious bursal disease virus-infected cells. *J Gen Virol.* (1995) 76 (Pt 2):437–43. doi: 10.1099/0022-1317-76-2-437

9. Azad AA, McKern NM, Macreadie IG, Failla P, Heine HG, Chapman A, et al. Physicochemical and immunological characterization of recombinant host-protective antigen (VP2) of infectious bursal disease virus. *Vaccine.* (1991) 9:715–22. doi: 10.1016/0264-410X(91)90286-F

10. Fahey KJ, O'Donnell IJ, Bagust TJ. Antibody to the 32K structural protein of infectious bursal disease virus neutralizes viral infectivity *in vitro* and confers protection on young chickens. *J Gen Virol.* (1985) 66(Pt 12):2693–702. doi: 10.1099/0022-1317-66-12-2693

11. Spies U, Müller H. Demonstration of enzyme activities required for cap structure formation in infectious bursal disease virus, a member of the birnavirus group. *J Gen Virol.* (1990) 71(Pt 4):977–81. doi: 10.1099/0022-1317-71-4-977

12. Müller H, Mundt E, Eterradossi N, Islam MR. Current status of vaccines against infectious bursal disease. *Avian Pathol.* (2012) 41:133–9. doi: 10.1080/03079457.2012.661403

13. Niewiesk S. Maternal antibodies: clinical significance, mechanism of interference with immune responses, and possible vaccination strategies. *Front Immunol.* (2014) 5:446–446. doi: 10.3389/fimmu.2014.00446

14. Block H, Meyer-Block K, Rebeski DE, Scharr H, de Wit S, Rohn K, et al. A field study on the significance of vaccination against infectious bursal disease virus (IBDV) at the optimal time point in broiler flocks with maternally derived IBDV antibodies. *Avian Pathol.* (2007) 36:401–9. doi: 10.1080/03079450701589175

15. Rautenschlein S, Kraemer C, Vanmarcke J, Montiel E. Protective efficacy of intermediate and intermediate plus infectious bursal disease virus (IBDV) vaccines against very virulent IBDV in commercial broilers. *Avian Dis.* (2005) 49:231–7. doi: 10.1637/7310-112204R

16. Tsukamoto K, Tanimura N, Kakita S, Ota K, Mase M, Imai K, et al. Efficacy of three live vaccines against highly virulent infectious bursal disease virus in chickens with or without maternal antibodies. *Avian Dis.* (1995) 39:218–29. doi: 10.2307/1591863

17. Kumar K, Singh KC, Prasad CB. Immune responses to intermediate strain IBD vaccine at different levels of maternal antibody in broiler chickens. *Trop Anim Health Prod.* (2000) 32:357. doi: 10.1023/A:1005225501513

18. Chauhan V, Rungta T, Goyal K, Singh MP. Designing a multi-epitope based vaccine to combat Kaposi Sarcoma utilizing immunoinformatics approach. *Sci Rep.* (2019) 9:2517. doi: 10.1038/s41598-019-39299-8

19. Lu I-N, Farinelle S, Sausy A, Muller CP. Identification of a CD4 T-cell epitope in the hemagglutinin stalk domain of pandemic H1N1 influenza virus and its antigen-driven TCR usage signature in BALB/c mice. *Cell Mol Immunol.* (2017) 14:511–20. doi: 10.1038/cmi.2016.20

20. Lennerz V, Gross S, Gallerani E, Sessa C, Mach N, Boehm S, et al. Immunologic response to the survivin-derived multi-epitope vaccine EMD640744 in patients with advanced solid tumors. *Cancer Immunol Immunother.* (2014) 63:381–94. doi: 10.1007/s00262-013-1516-5

21. Jiang P, Cai Y, Chen J, Ye X, Mao S, Zhu S, et al. Evaluation of tandem *Chlamydia trachomatis* MOMP multi-epitopes vaccine in BALB/c mice model. *Vaccine.* (2017) 35:3096–103. doi: 10.1016/j.vaccine.2017.04.031

22. Zhu S, Feng Y, Rao P, Xue X, Chen S, Li W, et al. Hepatitis B virus surface antigen as delivery vector can enhance *Chlamydia trachomatis* MOMP multi-epitope immune response in mice. *Appl Microbiol Biotechnol.* (2014) 98:4107–17. doi: 10.1007/s00253-014-5517-x

23. Saadi M, Karkhah A, Nouri HR. Development of a multi-epitope peptide vaccine inducing robust T cell responses against brucellosis using immunoinformatics based approaches. *Infect Genet Evol.* (2017) 51:227–34. doi: 10.1016/j.meegid.2017.04.009

24. Lu C, Meng S, Jin Y, Zhang W, Li Z, Wang F, et al. A novel multi-epitope vaccine from MMSA-1 and DKK 1 for multiple myeloma immunotherapy. *Br J Haematol.* (2017) 178:413–26. doi: 10.1111/bjh.14686

25. Lin X, Chen S, Xue X, Lu L, Zhu S, Li W, et al. Chimerically fused antigen rich of overlapped epitopes from latent membrane protein 2 (LMP2) of Epstein–Barr virus as a potential vaccine and diagnostic agent. *Cell Mol Immunol.* (2016) 13:492–501. doi: 10.1038/cmi.2015.29

26. Yin D, Li L, Song X, Li H, Wang J, Ju W, et al. A novel multi-epitope recombined protein for diagnosis of human brucellosis. *BMC Infect Dis.* (2016) 16:219. doi: 10.1186/s12879-016-1552-9

27. Cherryholmes GA, Stanton SE, Disis MLJV. Current methods of epitope identification for cancer vaccine design. *Vaccine.* (2015) 33:7408–14. doi: 10.1016/j.vaccine.2015.06.116

28. Zhang L. Multi-epitope vaccines: a promising strategy against tumors and viral infections. *Cell Mol Immunol.* (2018) 15:182–4. doi: 10.1038/cmi.2017.92

29. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics.* (2007) 8:4. doi: 10.1186/1471-2105-8-4

30. Mugunthan SP, Harish MC. Multi-epitope-based vaccine designed by targeting cytoadherence proteins of *Mycoplasma gallisepticum*. *ACS Omega.* (2021) 6:13742–55. doi: 10.1021/acsomega.1c01032

31. Omony JB, Wanyana A, Mugimba KK, Kirunda H, Nakavuma JL, Otim-Onapa M, et al. Epitope peptide-based predication and other functional regions of antigenic F and HN proteins of waterfowl and poultry avian avulavirus serotype-1 isolates from Uganda. *Front Vet Sci.* (2021) 8:610375. doi: 10.3389/fvets.2021.610375

32. Tan L, Liao Y, Fan J, Zhang Y, Mao X, Sun Y, et al. Prediction and identification of novel IBV S1 protein derived CTL epitopes in chicken. *Vaccine.* (2016) 34:380–6. doi: 10.1016/j.vaccine.2015.11.042

33. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M, et al. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics.* (2007) 8:424. doi: 10.1186/1471-2105-8-424

34. Andreatta M, Nielsen MJB. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics.* (2016) 32:511–7. doi: 10.1093/bioinformatics/btv639

35. Adhikari UK, Rahman MM. Overlapping CD8+ and CD4+ T-cell epitopes identification for the progression of epitope-based peptide vaccine from nucleocapsid and glycoprotein of emerging Rift Valley fever virus using immunoinformatics approach. *Infect Genet Evol.* (2017) 56:75–91. doi: 10.1016/j.meegid.2017.10.022

36. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics.* (2007) 8:238. doi: 10.1186/1471-2105-8-238

37. Dhanda SK, Vir P, Raghava GPS. Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct.* (2013) 8:30. doi: 10.1186/1745-6150-8-30

38. Dhanda SK, Gupta S, Vir P, Raghava GP. Prediction of IL4 inducing peptides. *Clin Dev Immunol.* (2013) 2013:263952. doi: 10.1155/2013/263952

39. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins.* (2006) 65:40–8. doi: 10.1002/prot.21078

40. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A, et al. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics.* (2006) 7:153. doi: 10.1186/1471-2105-7-153

41. Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v2–a server for *in silico* prediction of allergens. *J Mol Model.* (2014) 20:2278. doi: 10.1007/s00894-014-2278-5

42. Dimitrov I, Flower DR, Doytchinova I. AllerTOP–a server for *in silico* prediction of allergens. *BMC Bioinformatics.* (2013) 14:S4. doi: 10.1186/1471-2105-14-S6-S4

43. Jacob CO, Leitner M, Zamir A, Salomon D, Arnon R. Priming immunization against cholera toxin and *E. coli* heat-labile toxin by a cholera toxin short peptide-beta-galactosidase hybrid synthesized in *E coli*. *EMBO J.* (1985) 4:3339–43. doi: 10.1002/j.1460-2075.1985.tb04086.x

44. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* (1999) 112:531–52. doi: 10.1385/1-59259-584-7:531

45. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics.* (2000) 16:404–5. doi: 10.1093/bioinformatics/16.4.404

46. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

47. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M, et al. ColabFold - making protein folding accessible to all. (2022) 19:679–82. doi: 10.1101/2021.08.15.456425

48. Heo L, Park H, Seok C. GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* (2013) 41(Web Server issue):W384–8. doi: 10.1093/nar/gkt458

49. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* (2007) 35(suppl_2):W407–10. doi: 10.1093/nar/gkm290

50. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. *Proteins.* (1993) 17:355–62. doi: 10.1002/prot.340170404

51. Yan Y, Tao H, He J, Huang SY. The HDOCK server for integrated protein–protein docking. *Nat Protoc.* (2020) 15:1829–52. doi: 10.1038/s41596-020-0312-x

52. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun.* (1995) 91:43–56. doi: 10.1016/0010-4655(95)00042-E

53. Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* (2005) 33(Web Server issue):W526–31. doi: 10.1093/nar/gki376

54. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C, et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.* (2015) 11:3696–713. doi: 10.1021/acs.jctc.5b00255

55. Rapin N, Lund O, Bernaschi M, Castiglione F. Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system. *PLoS ONE.* (2010) 5:e9862. doi: 10.1371/journal.pone.0009862

56. Vartak A, Sucheck SJ. Recent advances in subunit vaccine carriers. *Vaccines.* (2016) 4:12. doi: 10.3390/vaccines4020012

57. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T- and B-cell epitope prediction. *J Immunol Res.* (2017) 2017:2680160. doi: 10.1155/2017/2680160

58. Khan MT, Islam MJ, Parihar A, Islam R, Jerin TJ, Dhote R, et al. Immunoinformatics and molecular modeling approach to design universal multi-epitope vaccine for SARS-CoV-2. *Inform Med Unlocked.* (2021) 24:100578. doi: 10.1016/j.imu.2021.100578

59. Singh H, Jakhar R, Sehrawat N. Designing spike protein (S-Protein) based multi-epitope peptide vaccine against SARS COVID-19 by immunoinformatics. *Heliyon.* (2020) 6:e05528. doi: 10.1016/j.heliyon.2020.e05528

60. Chukwudozie OS, Gray CM, Fagbayi TA, Chukwuanukwu RC, Oyebanji VO, Bankole TT, et al. Immuno-informatics design of a multimeric epitope peptide based vaccine targeting SARS-CoV-2 spike glycoprotein. *PLoS ONE.* (2021) 16:e0248061. doi: 10.1371/journal.pone.0248061

61. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. In: JM Walker, editor. *The Proteomics Protocols Handbook*. Totowa, NJ: Humana Press (2005), p. 571–607. doi: 10.1385/1-59259-890-0:571

62. Boehme KW, Compton T. Innate sensing of viruses by toll-like receptors. *J Virol.* (2004) 78:7867–73. doi: 10.1128/JVI.78.15.7867-7873.2004

63. Choi Y, Bowman JW, Jung JU. Autophagy during viral infection - a double-edged sword. *Nat Rev Microbiol.* (2018) 16:341–54. doi: 10.1038/s41579-018-0003-6

64. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* (2005) 33(Database issue):D266–8. doi: 10.1093/nar/gki001

65. Laskowski RA, Jabłońska J, Pravda L, Vareková RS, Thornton JM. PDBsum: structural summaries of PDB entries. *Protein Sci.* (2018) 27:129–34. doi: 10.1002/pro.3289

66. Nicholson D, Nicholson LB. A simple immune system simulation reveals optimal movement and cell density parameters for successful target clearance. *Immunology.* (2008) 123:519–27. doi: 10.1111/j.1365-2567.2007.02721.x

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 19 January 2023
DOI 10.3389/fvets.2022.1085112

Check for updates

OPEN ACCESS

EDITED BY
Rajesh Kumar Pathak,
Chung-Ang University, Republic of Korea

REVIEWED BY
Joycee Jogi,
Nanaji Deshmukh Veterinary Science University, India
Rohit Shukla,
Jaypee University of Information Technology, India
Ravendra Chauhan,
Rhodes University, South Africa

*CORRESPONDENCE
Mostafa Ghaderi-Zefrehei
✉ mosmos741@yahoo.com;
✉ mghaderi@yu.ac.ir
Jacqueline Smith
✉ jacqueline.smith@roslin.ed.ac.uk

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

RECEIVED 31 October 2022
ACCEPTED 28 December 2022
PUBLISHED 19 January 2023

CITATION
Davoodi P, Ghaderi-Zefrehei M, Dolatabady MM, Razmkabir M, Kianpour S, Esfahani EN and Smith J (2023) *In silico* investigation of uncoupling protein function in avian genomes. *Front. Vet. Sci.* 9:1085112. doi: 10.3389/fvets.2022.1085112

In silico investigation of uncoupling protein function in avian genomes

Peymaneh Davoodi [1], Mostafa Ghaderi-Zefrehei [2]*, Mustafa Muhaghegh Dolatabady [2], Mohammad Razmkabir [3], Somayeh Kianpour [1], Effat Nasre Esfahani [4] and Jacqueline Smith [5]*

[1]Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran, [2]Department of Animal Science, Faculty of Agriculture, Yasouj University, Yasouj, Iran, [3]Department of Animal Science, Faculty of Agriculture, University of Kurdistan, Sanandaj, Iran, [4]Department of Agriculture, Payam Noor University Tehran, Tehran, Iran, [5]The Roslin Institute and Royal (Dick) School of Veterinary Studies R(D)SVS, University of Edinburgh, Edinburgh, United Kingdom

Introduction: The uncoupling proteins (*UCPs*) are involved in lipid metabolism and belong to a family of mitochondrial anionic transporters. In poultry, only one *UCP* homologue has been identified and experimentally shown to be associated with growth, feed conversion ratio, and abdominal fat according to its predominant expression in bird muscles. In endotherm birds, cell metabolic efficiency can be tuned by the rate of mitochondrial coupling. Thus, *avUCP* may be a key contributor to controlling metabolic rate during particular environmental changes.

Methods: This study aimed to perform a set of *in-silico* investigations primarily focused on the structural, biological, and biomimetic functions of *avUCP*. Thereby, using *in silico* genome analyses among 8 avian species (chicken, turkey, swallow, manakin, sparrow, wagtail, pigeon, and mallard) and a series of bioinformatic approaches, we provide phylogenetic inference and comparative genomics of *avUCP*s and investigate whether sequence variation can alter coding sequence characteristics, the protein structure, and its biological features. Complementarily, a combination of literature mining and prediction approaches was also applied to predict the gene networks of *avUCP* to identify genes, pathways, and biological crosstalk associated with *avUCP* function.

Results: The results showed the evolutionary alteration of *UCP* proteins in different avian species. Uncoupling proteins in avian species are highly conserved trans membrane proteins as seen by sequence alignment, physio-chemical parameters, and predicted protein structures. Taken together, *avUCP* has the potential to be considered a functional marker for the identification of cell metabolic state, thermogenesis, and oxidative stress caused by cold, heat, fasting, transfer, and other chemical stimuli stresses in birds. It can also be deduced that *avUCP*, in migrant or domestic birds, may increase heat stress resistance by reducing fatty acid transport/b-oxidation and thermoregulation alongside antioxidant defense mechanisms. The predicted gene network for *avUCP* highlighted a cluster of 21 genes involved in response to stress and 28 genes related to lipid metabolism and the proton buffering system. Finally, among 11 enriched pathways, crosstalk of 5 signaling pathways including MAPK, adipocytokine, mTOR, insulin, ErbB, and GnRH was predicted, indicating a possible combination of positive or negative feedback among pathways to regulate *avUCP* functions.

Discussion: Genetic selection for fast-growing commercial poultry has unintentionally increased susceptibility to many kinds of oxidative stress, and so *avUCP* could be considered as a potential candidate gene for balancing energy expenditure and reactive oxygen species production, especially in breeding programs. In conclusion, *avUCP* can be introduced as a pleiotropic gene that requires the contribution of regulatory genes, hormones, pathways, and genetic crosstalk to allow its finely-tuned function.

KEYWORDS

avian uncoupling protein, nucleotide coding sequence, protein structure prediction, codon usage, pathway crosstalk

1. Introduction

Generalized homeostasis of energy expenditure and energy intake is essential for the best selection criteria in poultry, however, the regulatory mechanisms connecting feed intake, growth, and energy balance are still confusing. A growing body of literature implies avian Uncoupling Protein (*avUCP*) plays a key role in cell metabolism and adaptive thermogenesis. It thus may have the potential to be considered the missing link in the chain of whole-body energy homeostasis in chickens. The *avUCP* is a homolog with more than 70% protein sequence similarity with mammalian *UCP3* and *UCP2*, harboring two conserved regions of the mitochondrial carrier and ADP/ATP transporter translocase (1, 2). The *avUCP* gene was first identified in 2001 after screening a hummingbird skeletal muscle cDNA library (3). Subsequent studies in different avian species revealed a predominant expression of *avUCP* mRNA in skeletal muscle in chicken (*Gallus gallus*), king penguin (*Aptenodytes patagonicus*), and hummingbird (*Eupetomena macroura*) (1, 4–7). The interconnection of growth, oxidative stress, reproductive state, immunity, and feather coloration processes and their efficacy on thermo-regulation have been suggested by several studies (8–12).

For a long time, shivering thermogenesis has been known to be the main thermogenic mechanism in avian species (13, 14), although the evidence for the existence of adaptive mechanisms of heat production and non-shivering thermogenesis are currently growing (1). This thermogenesis mechanism can be boosted by increasing oxidative metabolic capacity along with the uncoupling of aerobic metabolism from ATP production. Also, previous studies support the involvement of *avUCP* in avian energy expenditure and adaptive thermogenesis (1, 15–18). Taouis et al. showed that early thermal conditioning in broiler chicks can instantly reduce body temperature and *avUCP* expression in the pectoral muscle, which may potentially improve the resistance to heat stress in broilers (15). In contrast, another study showed that diet-induced thermogenesis had no control over feed intake in layers and broilers and the expression of *avUCP* was not influenced by layer and broiler genotypes. Consequently, these findings led to the rejection of the hypothesis of the involvement of *avUCP* in diet-induced thermogenesis (19). A study in ducklings has verified that, in proportion to the degree of cold, the increase in metabolic heat production occurs in parallel with the upregulation of *avUCP* and higher mitochondrial oxidative phosphorylation, while no change in mitochondrial membrane conductance capacity occurred (20).

avUCP has also been suggested to be involved in cell metabolism. Several studies have identified polymorphism of the *avUCP* gene that is associated with fat metabolism, growth, feed intake, and exposure to abiotic stress conditions (2, 21, 22). Additionally, it was also revealed that the upregulation of *avUCP* can result in the down-regulation of reactive oxygen species (ROS) production in the skeletal muscle of fasted chickens (23). On the other hand, the *avUCP* expressed in glycolytic muscle fibers may be a passive transporter of pyruvate for ensuring a sustained balance between glycolysis and oxidative phosphorylation (24). Conversely, it was demonstrated that heat

Abbreviations: ADP, Adenosine diphosphate; AI, Aliphatic index; AMPK, AMP-activated protein kinase; ATP, Adenosine triphosphate; CAI, Codon Adaptation Index; cDNA, Complementary DNA; CDS, Coding sequence; CS, Citrate synthase; CU, Codon usage; FDR, False discovery rate; GDP, Guanosine diphosphate; GO, Gene ontology; HK1, Hexokinase 1; HK2, Hexokinase 2; II, Instability index; IL-6, Interleukin 6; LDHA, Lactate dehydrogenase A; LDHB, Lactate dehydrogenase B; mRNA, Messenger RNA; PDHX, Pyruvate dehydrogenase complex component X; PFK, Phosphofructokinase; PGC-1α, Peroxisome proliferator-activated receptor γ coactivator-1α; P_i, Inorganic phosphate; PK, Protein kinase; PPARGC1A, PPARG coactivator 1 alpha; PPARs, Peroxisome proliferator-activated receptors; ROS, Reactive oxygen species; RSCU, Relative Synonymous Codon Usage; SDHA, Succinate dehydrogenase complex flavoprotein subunit A; SDHB, Succinate dehydrogenase complex iron sulfur subunit B; SLC25A4, Solute Carrier Family 25 Member 4; TNFα, Tumor necrosis factor alpha; TORC1, TOR complex 1; UCP, Uncoupling Protein.

stress stimulates mitochondrial superoxide production in broiler skeletal muscle through the downregulation of uncoupling protein (25, 26). Another study showed that the expression of members of the beta-oxidation pathway and mitochondrial fatty acid transport were upregulated upon heat stress. However, the expression of *avUCP* did not control ROS production in heat-stressed chickens (18). Heat stress, by causing oxidative stress, impairs mitochondrial function by decreasing *avUCP* expression (27) which can further impair meat quality and increases glycolysis and intramuscular fat deposition (2, 28, 29). Additionally, the upregulation of *avUCP* and *avPGC-1α* together can help to reduce ROS accumulation and lipid oxidation in the skeletal muscles of birds (28, 30). Several distinct studies of *avUCP*, have demonstrated the regulation of *avUCP* expression and regulation of its putative function. Accordingly, thyroid hormones were reported to increase thermogenic capacity in the avian muscle and liver (31, 32). Moreover, uncoupling of sarco- endoplasmic reticulum calcium ATPase pump activity in muscle (33) and regulation of glycolysis are involved in controlling thermogenic processes in avian species. Some studies have implied that thermogenesis is controlled by thyroid hormone affecting *PPARGC1A* and *SLC25A4* gene expression in chickens (34, 35). However, triiodothyronine (T3) is reported to have a biphasic effect on *avUCP* expression (32). Another study investigating variations in *avUCP* expression, thyroid hormone metabolism, and heat production during cold exposure has reported a significant increase in body temperature, *avUCP* expression, *T3* level, renal outer-ring deiodination activity, and also increased thyroxine (*T4*) level, and hepatic inner-ring deiodination activity. Meanwhile, no significant differences in body weight and feed intake were reported in comparison with chickens reared in normal temperatures (36).

Moreover, it has been implied that *avUCP* gene expression is down-regulated by leptin hormone and up-regulated by pro-inflammatory cytokines *IL-6* and *TNFα* through modulation of *avUCP*-related transcription factors (*PPAR*s and *PGC-1α*) (32). Two-fold over-expression in gastrocnemius muscle, significant down-regulation, and no significant change were reported in *avUCP* mRNA expression through injection of thyroid hormone, methimazole, and insulin respectively (16). Additionally, selenium deficiency in broilers can cause a reduction in *avUCP* mRNA levels that results in oxidative stress, inflammation, and glyco-metabolism disorders (37).

Furthermore, in fat chickens with a higher fat diet, *avUCP* was significantly up-regulated, which could be correlated with the particular need for antioxidant pathways in muscle (38). Previous research has provided some evidence for the involvement of the beta-adrenergic system, *PPAR* transcription factors, and the AMP-activated protein kinase (*AMPK*) to control the expression of *avUCP* (39). Furthermore, oral use of D-aspartate resulted in a reduction in body temperature through the decline in *avUCP* mRNA expression in the breast muscle, which may be involved in reduced mitochondrial proton leaks and heat production (40).

There is also some evidence Oleuropein can also affect *avUCP* expression as well as genes related to mitochondrial oxidative phosphorylation and induce mitochondrial biogenesis in avian muscle cells. Oleuropeins can suppress mitochondrial superoxide production, through up-regulation of *avUCP* and manganese superoxide dismutase (41, 42). Therefore, the orexin system in avian muscle cells can regulate mitochondrial dynamics without affecting ATP synthesis (43). Evidence has also been presented that retinoic acid can activate the thermogenic function of *avUCP* in birds (44). Interestingly, "avian" is reported to be the only vertebrate lineage having just one *UCP* gene (45). Thus, the avian uncoupling protein seems to provide a unique opportunity to explore the functional activity and regulation patterns of *UCP*. We aimed, therefore, to investigate *avUCP*s in eight different avian species (chicken, turkey, swallow, manakin, sparrow, wagtail, pigeon, and mallard) through a wide range of comparative bioinformatics analyses to better understand the details of *avUCP*s, from their coding sequences to their functional consequences.

2. Methods

2.1. Coding sequence analysis

The nucleotide coding sequences and amino acid sequences of avian uncoupling proteins were downloaded from a dataset contained at https://figshare.com/. Sequence alignment of coding sequences (CDS) for eight avian species including chicken, turkey, swallow, manakin, sparrow, wagtail, pigeon, and mallard was conducted for determining the number of conserved, variable, parsimonious, and singleton sites in *avUCPs*. Moreover, nucleotide composition, GC content, codon frequency, and relative synonymous codon usage were obtained using MEGA11 software (46). The Codon Adaptation Index (CAI) for each of the studied avian species was estimated using the Markov model with 500 replications over the *avUCP* DNA sequence. As the reference set to calculate the CAI is important for interpretation, the codon usage table for each species (if it exists), from the codon usage database on the *CAIcal* server (http://www.kazusa.or.jp/codon/) was therefore utilized (47). The Relative Synonymous Codon Usage (RSCU) was calculated as follows:

$$RSCU_{i,j} = \frac{n_i x_{i,j}}{\sum_{j=1}^{n_i} x_{i,j}} \tag{1}$$

Where x_i is the number of times the *ith* codon has been favored to be used for an amino acid, and n represents the number of synonymous codons for that amino acid.

2.2. Protein sequence analysis

Amino acid (a.a) composition, physio-chemical parameters and phylogenetic analysis of eight avian protein sequences were performed in QIAGEN *CLC* Genomics Workbench (RRID: SCR_011853) (48). Physio-chemical parameters including molecular weight, isoelectric point, extinction coefficient, instability index, aliphatic index, and grand average of hydropathicity, were determined for each *avUCP* protein sequence. The phylogenetic analyses were performed using the Neighbor-Joining method, Jukes-Cantor protein distance measure, and bootstrapping over 307 a.a of protein sequences of eight avian species in *CLC* Genomics Workbench (48). Entropy analysis was then carried out using *BioEdit* (49) to further determine variable and conserved sites and finally, by using the *Skylign* online tool, a positional logo of amino acid variability was constructed (50).

2.3. Protein structure prediction

The secondary structures of *avUCPs* were predicted by the *SOPMA* predictor (51). All *avUCP* sequences were submitted to the *Phyre2* web portal (http://www.sbg.bio.ic.ac.uk/$\sim$phyre2/html/page.cgi?id=index) as a batch file for tertiary structure prediction. After that, the *avUCPs* were modeled through four stages including homology detection, fold library scanning to predict secondary structure, loop modeling, and sidechain fitting (52). Structural evaluation and qualification were then performed using the *Swiss-Model* online tool (53).

2.4. Sequence-based gene ontology prediction

PredictProtein was used to predict Gene Ontology (GO) terms of cellular components, molecular function, and biological process for *avUCP* protein sequences (54). In this process, the distance between the input protein sequence and the closest annotated protein represents the reliability of GO prediction (54).

2.5. Interactive network prediction and gene-based enrichment analysis

By identifying genes related to *avUCP* from the literature, a list of genes was extracted according to Davoodi and Ehsani (55) for protein-protein network prediction. The list of most related-genes was provided through the retrospective review of previous studies on *avUCP* (5, 42, 56–71). Biomolecular network prediction and gene set enrichment analysis of networked genes were performed in *Cytoscape* (72) using *STRING* v11.5 (73, 74).

2.6. Pathway crosstalk prediction

Crosstalk prediction was applied using *XtalkDB* by querying pathways enriched for *avUCP* to predict which pairs of signaling pathways may interact to reach a conclusive understanding of biological pathways involved in the regulation of *avUCP* functions from a global view (75).

2.7. Datasets

The nucleotide coding sequences (CDS) and protein sequences of avian uncoupling protein (*avUCP*) from eight different avian species (chicken, turkey, swallow, manakin, sparrow, wagtail, pigeon, and mallard) were retrieved in FASTA format from the NCBI database and used for *in silico* analyses (Table 1).

2.8. Model of analysis

This research had an integrative pipeline but no unique statistical model. All parts of the pipeline are explained previously in each section.

3. Results

3.1. Coding sequence analysis

3.1.1. Nucleotide composition

The CDS sequences were analyzed for nucleotide composition, GC content, conserved, variable, parsimony informative, and singleton sites. The CDS length in all selected birds consisted of 924 nucleotides. The number of conserved, variable, parsimony informative, and singleton sites were revealed as 684, 240, 145, and 95, respectively. Divergence details of GC content among the eight avian species are shown in Table 2. The "C" content in the coding sequences of *avUCP* in wagtail (2.5), mallard (1.0), sparrow (0.8), and manakin (0.1) was higher than that of "G," however, the "G" content in turkey (1.8), chicken (1.2), swallow (0.6), and pigeon (0.3) was higher than that of the "C" content.

3.1.2. Codon usage analysis

The codon usage (CU) and relative synonymous codon usage (RSCU) values of *avUCP* coding sequence were calculated, then CU and RSCU patterns were obtained from the eight avian

TABLE 1 General information for *avUCP* genes in eight avian species extracted from the NCBI database.

Species	Gene ID	Chromosome	Exon number	Transcript ID
Gallus gallus (chicken)	373896	1	6	NM_204107.2
Anas platyrhynchos (mallard)	101794508	1	6	XM_005025525.4
Chiroxiphia lanceolata (lance-tailed Manakin)	116781978	2	7	XM_032677900.1
Columba livia (rock pigeon)	102092157	Unknown	8	XM_021285112.1
Passer montanus (eurasian tree sparrow)	120496512	Unknown	7	XM_039697035.1
Hirundo rustica (barn swallow)	120765208	2	7	XM_040089804.1
Meleagris gallopavo (turkey)	100303663	1	6	NM_001303164.1
Motacilla alba alba (white wagtail)	119699879	1	7	XM_038133918.1

TABLE 2 GC content (%) of *avUCP* in eight different avian species.

	C	G	C-1	G-1	C-2	G-2	C-3	G-3	GC	GC-1	GC-2	GC-3
Chicken	32.5	33.7	26.6	35.1	26.0	23.1	44.8	42.9	66.1	63.8	49.7	87.6
Mallard	35.0	34.0	26.6	37.7	26.9	23.7	51.3	40.6	68.9	66.4	51.0	91.9
Manakin	32.0	31.9	26.0	37.7	26.0	22.1	44.2	36.0	64.0	65.8	48.7	79.9
Pigeon	34.1	34.4	26.0	37.3	27.3	23.4	49.0	42.5	68.5	65.2	51.2	91.6
Sparrow	34.2	33.4	25.3	37.0	26.3	23.4	51.0	39.9	67.6	65.3	51.0	90.8
Swallow	33.8	34.4	26.0	37.0	26.6	24.4	48.7	41.9	68.2	65.8	52.2	90.5
Turkey	31.8	33.7	26.3	35.1	26.0	23.1	43.2	42.9	65.5	63.6	49.8	85.9
Wagtail	35.8	33.3	25.6	37.7	26.3	23.4	55.5	39.0	69.2	66.6	50.9	94.5

avUCPs. Generally, 64 combinations of 3-letter codons encode 20 different amino acids, thus showing codon redundancy. After excluding the three stop codons, 25 codons in pigeon, sparrow, turkey, wagtail, 24 codons in chicken, mallard, and swallow, and 23 codons in manakin *avUCP* were observed with an RSCU value higher than 1. Moreover, the numbers of unused codons (RSCU = 0) were as follows: chicken—13, mallard—14, manakin—8, pigeon—16, swallow—13, sparrow—15, turkey—11, and wagtail—18. The RSCU value of the codon CUG, which encodes leucine, was the highest in all selected avian species. By looking at codons with an RSCU >1 and examining their final bases, it was found that they ended, on average, with C (15), G (9), and roughly one U and A in the selected species. The highly preferred codons within *avUCP*, with their corresponding CU and RSCU values are presented in Table 3. As can be seen, all highly preferred codons, except the CAU (only in manakin), which encodes for histidine, end with a "C" or a "G." In addition, these 25 highly preferred codons are responsible for encoding around 76% (in manakin) to 89% (in wagtail) of the total protein sequences of *UCP*. Moreover, the expected codon adaptation index (CAI) for retrieved CDSs of chicken, mallard, manakin, pigeon, sparrow, swallow, turkey, and wagtail were 0.707, 0.691, 0.712, 0.697, 0.695, 0.692, 0.709, and 0.696, respectively.

3.2. Protein sequence analysis

3.2.1. Amino acid compositions

The results of the protein sequence analysis of the eight avUCPs revealed a sequence length of 307 amino acids and the amino acid compositions in different birds are represented in Figure 1. Alanine, leucine, valine, and glycine have been observed at high frequency in all *avUCP*s, while histidine—a positively charged, and tryptophan—an aromatic amino acid, have been detected at the lowest frequency in all *avUCPs*. In contrast to the slight variation in amino acid usage among the eight studied *avUCP* sequences, tyrosine was the only completely constant amino acid among the *avUCP*s in all eight birds.

3.2.2. Physio-chemical analysis

Physio-chemical parameters of uncoupling proteins including molecular weight, isoelectric point, aliphatic index, number of sulfur atoms, hydrophobicity, hydrophilicity, the percentage of negatively and positively charged amino acids, instability index, and grand average hydropathy of avian uncoupling protein in eight avian species are summarized in Table 4. The isoelectric point of *UCP*s ranged from 9.51 to 9.66. The lowest instability index (II) and the highest aliphatic index

TABLE 3 The codon usage and relative synonymous codon usage of highly preferred codons for *avUCP* in eight avian species.

Codon	Chicken		Mallard		Manakin		Pigeon		Swallow		Sparrow		Turkey		Wagtail	
	CU	RSCU	CU	RSCU	CU	RSCU	CU	RSCU	CU	RSCU	CU	RSCU	CU	RSCU	CU	RSCU
AGC(S)	11	3.5	11	3.7	7	2.8	9	3.4	11	3.3	9	3.2	10	3.2	10	3.5
CGG(R)	8	2.1	6	1.6	7	2.0	8	2.1	12	3.0	6	1.6	8	2.1	9	2.4
CGC(R)	7	1.8	9	2.5	9	2.6	11	2.9	8	2.0	12	3.1	4	1.0	10	2.6
CUG(L)	25	4.4	20	3.8	20	3.2	20	3.8	22	3.9	21	3.9	24	4.4	20	3.8
CUC(L)	8	1.4	9	1.7	12	2.0	11	2.1	11	1.9	10	1.9	8	1.5	11	2.1
GCC(A)	18	2.3	22	2.4	21	2.2	20	2.3	30	3.2	28	3.0	20	2.6	31	3.4
GGG(G)	14	2.2	16	2.3	13	1.9	14	1.9	12	1.7	11	1.6	13	2.0	7	1.0
GGC(G)	9	1.4	11	1.6	9	1.3	9	1.2	11	1.6	11	1.6	7	1.1	18	2.7
GUG(V)	27	3.5	21	2.6	20	2.7	21	2.7	22	3.1	23	3.1	27	3.5	21	2.6
GUC(V)	4	0.5	10	1.3	7	0.9	9	1.2	5	0.7	6	0.8	4	0.5	10	1.3
ACC(T)	10	1.7	10	2.0	10	2.0	11	1.6	10	1.9	10	1.9	11	1.9	9	1.7
ACG(T)	9	1.6	9	1.8	3	0.6	11	1.6	7	1.3	8	1.5	10	1.7	8	1.5
CCC(P)	13	2.9	13	2.6	9	2.4	12	3.2	10	2.7	10	2.5	13	2.9	11	2.8
AUC(I)	10	2.5	11	3.0	8	2.4	10	3.0	10	3.0	10	2.7	11	2.8	8	3.0
AAG(K)	7	1.6	7	1.6	10	1.8	8	1.8	8	1.8	8	1.8	7	1.6	9	2.0
AAC(N)	8	1.8	7	2.0	6	1.3	8	2.0	6	1.7	7	1.8	6	1.3	8	1.8
CAG(Q)	11	1.8	9	1.5	13	1.9	13	2.0	11	2.0	12	2.0	12	2.0	11	1.8
CAC(H)	1	2.0	2	2.0	1	2.0	1	2.0	1	2.0	0	0.0	1	2.0	1	2.0
CAU(H)	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
GAG(E)	10	2.0	8	1.8	8	1.8	8	1.8	9	2.0	9	2.0	10	2.0	9	2.0
GAC(D)	7	1.4	11	2.0	6	1.1	10	1.8	9	1.6	10	1.8	8	1.6	10	1.8
UAC(Y)	9	1.6	11	2.0	9	1.6	8	1.5	10	1.8	11	2.0	9	1.6	11	2.0
UGC(C)	7	1.8	7	1.8	7	1.8	7	1.8	7	1.8	8	1.8	7	1.8	7	1.8
UUC(F)	11	2.0	9	1.8	9	1.8	10	1.8	7	1.4	9	1.8	10	1.8	9	1.8
UGG(W)	2	1.0	3	1.0	2	1.0	2	1.0	2	1.0	2	1.0	2	1.0	3	1.0
AUG(M)	7	1.0	6	1.0	7	1.0	6	1.0	10	1.0	11	1.0	8	1.0	11	1.0

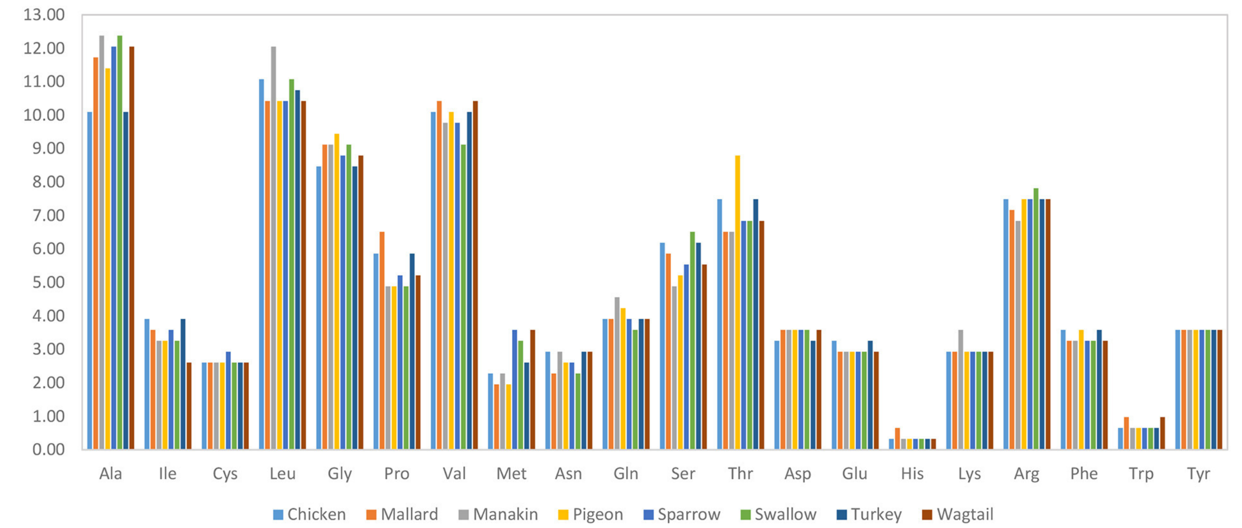


FIGURE 1
Amino acid composition of *avUCP* in different avian species.

TABLE 4 Physio-chemical parameters of uncoupling proteins from eight avian species.

Species	MW (kDa)	IP	AI	S	H-phobic	H-philic	−R	+R	II	GRAVY
Chicken	33.13	9.58	97.82	15	0.56	0.27	0.065	0.104	41.04	0.200
Mallard	32.81	9.51	96.58	14	0.58	0.25	0.065	0.101	39.74	0.209
Manakin	32.85	9.56	100.42	15	0.58	0.25	0.065	0.104	34.72	0.237
Pigeon	32.85	9.58	94.04	14	0.56	0.27	0.065	0.104	35.03	0.174
Sparrow	32.96	9.54	95.02	20	0.57	0.25	0.065	0.104	39.73	0.236
Swallow	32.84	9.66	94.72	18	0.57	0.25	0.065	0.107	38.59	0.215
Turkey	33.15	9.58	96.55	16	0.56	0.27	0.065	0.104	39.73	0.194
Wagtail	33.13	9.58	93.09	19	0.57	0.25	0.065	0.104	37.28	0.197

MW, Molecular weight; kDa, kilodalton; IP, Isoelectric point; AI, Aliphatic index; S, number of sulfur atom; H-phobic, Hydrophobic%; H-philic, Hydrophilic; −R, negatively charged; +R, positively charged; II, instability index; GRAVY, grand average hydropathy.

(AI) were observed in *UCP* of manakin. The molecular weight and the overall negatively/positively charged amino acids were nearly similar in all *avUCPs* in the current study.

The atomic sulfur count varied from 14 (mallard, pigeon) to 20 (sparrow). Sulfur can be found in cysteine and methionine amino acids. Eight cysteine residues were observed in seven of the *avUCPs* but the protein sequence of sparrow contained nine cysteine residues. Furthermore, the hydrophobic methionine was variable among *avUCPs*, which could be a source of variability in atomic sulfur count among *avUCP* sequences in the studied species.

A protein with an II smaller than 40 is considered stable, and proteins with an II above 40 can be considered somehow unstable (76, 77) and in the current study, the highest II was observed in *UCP* of chicken. Also, protein sequences with a GRAVY index above 0 are more likely to be hydrophobic and all studied proteins were revealed to have more hydrophobic regions.

3.2.3. Entropy analysis

For more evaluation of the status of amino acids in the *avUCP* protein sequences, entropy measures for each position were estimated using *BioEdit* (49) and visualized by the Shannon entropy plot, as shown in Figure 2. The estimated entropies of aligned sequences ranged from 0 to 1.56, the average entropy was estimated as 0.133 and a total of 274 positions displayed an entropy of 0. Seven positions revealed entropy values higher than 1 (147, 151, 267, 306, 299, 150, and 307). Among these, 147th and 151st positions showed the highest entropy of 1.56

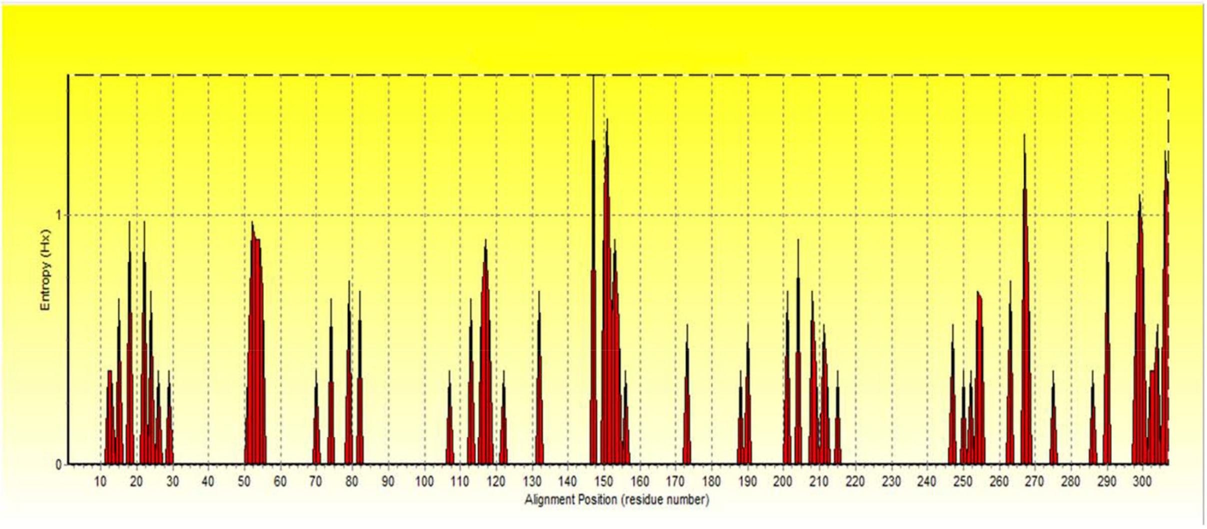


FIGURE 2
The Shannon entropy plot for alignment of proteins.

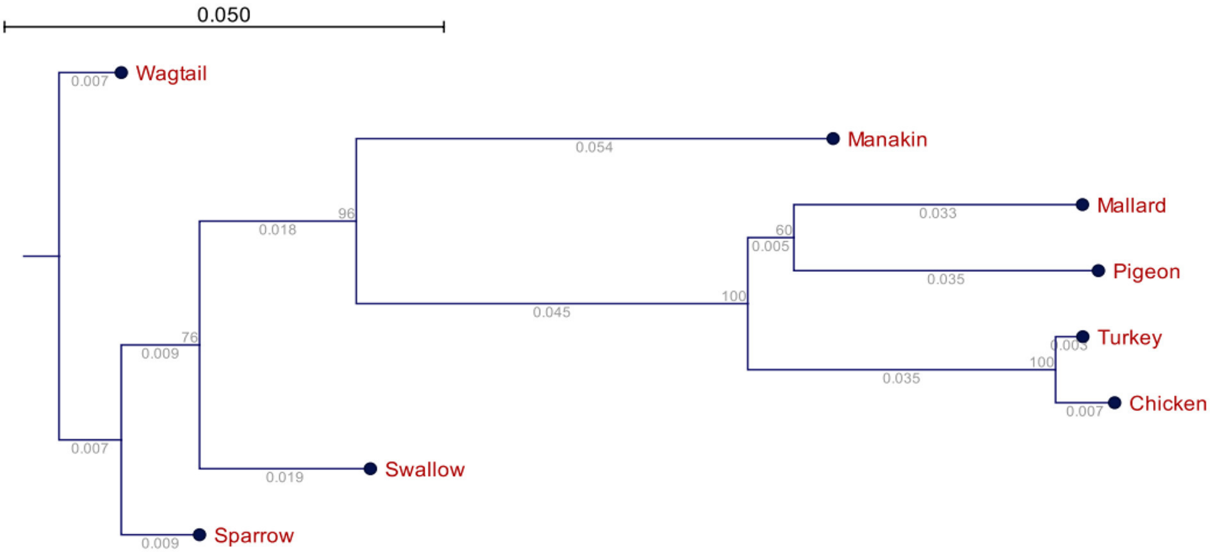


FIGURE 3
Phylogenetic relationships of the *avUCPs* in eight avian species (the numbers below the lines represent the branch size, and the numbers above the lines represent the bootstrap value).

and 1.39, respectively. Also, eight regions with a length of more than 10 amino acids (1→ 11, 30→ 50, 56→ 70, 83→ 107, 133→ 146, 157→ 172, 174→ 187, 216→ 246) have represented entropy of zero, indicating totally conserved regions in the examined *avUCP* protein sequences. Moreover, any region with more than 10 consecutive amino acids with an average entropy of >1 was not observed. To provide a compact representation of the most variable sites with the highest entropy in *avUCP* sequences among these eight species, the *Skylign* online tool (50) was used to make a positional logo showing amino acid variability (Figure 3).

3.2.3.1. Phylogeny analysis

The phylogenetic relationship of the *avUCPs* in the eight examined avian species is depicted in Figure 4. A neighbor-joining tree was derived from the multiple protein sequence alignment. Phylogenetically, the longest distance was detected between chicken and wagtail, in contrast chicken and turkey

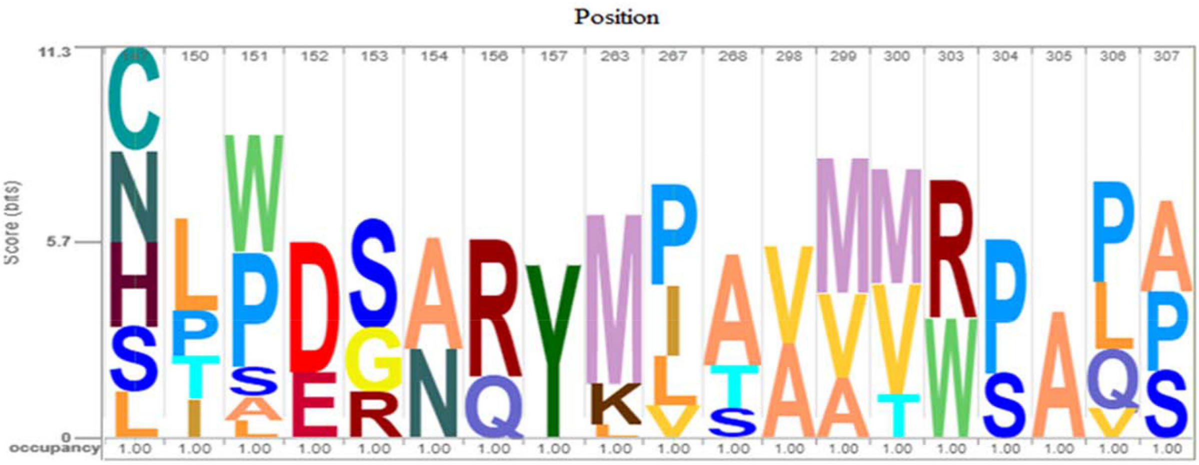


FIGURE 4
Logo of *avUCP* protein alignments (most variable sites with high entropy) among eight avian species. The height of the bar of letters (amino acids) displays the conservation at that position and the height of each letter within a bar is determined based on the frequency of that letter in that position.

showed the closest phylogenetic relationship which is in agreement with comparative results of amino acid component, codon usage pattern and physio-chemical parameters among *avUCP*s of mentioned birds.

3.3. Protein structure prediction

SOPMA was used to determine the percentage of α-helix, β-sheets, turns, and random coils to predict the secondary structure of the selected *avUCP* sequences through a neural network approach (51). The schematic predicted secondary structures of *avUCP*s in the eight avian species are shown in Figure 5. Because the secondary and tertiary structures of the protein are completely influenced by its primary structure, any differences in amino acid sequences can potentially modify the secondary and tertiary structures. Within the eight *avUCP* proteins, slightly different percentages of α-helical and β-turn formations were derived. The average contribution of alpha-helices, extended-strands, beta-turns, and random coils were calculated as 46.60 ± 1.04, 15.87 ± 0.58, 7.42 ± 0.65, and 30.08 ± 0.95% among the eight *avUCP* proteins, respectively. The level of alpha-helical structure was determined higher than other secondary structures.

The protein structure was predicted by three different software programs: phyre2, Predict Protein and SOPMA, with the final structural evaluation performed by SWISS_MODEL for checking the clashing score, favoured Ramachandran residues, and rotamer outliers. By using the *Phyre2web* portal (52), through the homology detection method, three-dimensional structures of the 8 *avUCP*s alongside potential extracellular, cytoplasmic, and trans membrane helices were predicted. In the protein structure, 6 trans membrane, 4 extracellular, and 3 cytosolic regions were predicted for the *avUCP*s. The two C-terminal and N-terminal segments were considered extracellular regions. Moreover, both hydrophobic and hydrophilic regions can be seen throughout the *avUCP* sequence. Accordingly, six areas with positive hydropathy are likely to represent transmembrane helices, which indicates *avUCP* can function as a trans membrane protein (Figure 6A). Meanwhile, areas with negative hydropathy show that these regions can form the extracellular part of *avUCP* (Figure 6A).

Ultimately, among the eight *avUCP*s, the predicted structure with the highest sequence identity (71%), alignment coverage (94%), interface similarity (51%), and confidence (100%) is illustrated in Figures 6A–C. The nuclear magnetic resonance molecular fragment replacement approach was applied to re-specify protein structure using the *Swiss-Model* web tool. The local and global model quality was then specified from Ramachandran analysis for the predicted tertiary structure of *avUCP* (Figure 6D). The final structure evaluation resulted in a clash score (the number of serious clashes per 1,000 atoms) of 90%, Ramachandran residues favored−88.46%, Ramachandran outliers−3.50%, rotamer outliers−8.09%, 0.01 bad angles, and two C-beta deviations.

3.4. Sequence-based gene ontology prediction

Gene Ontology (GO) terms associated with *avUCP* were predicted using deep learning embedding through the *PredictProtein* online tool (54). For this reason, the distance between the input *avUCP* protein and the closest annotated

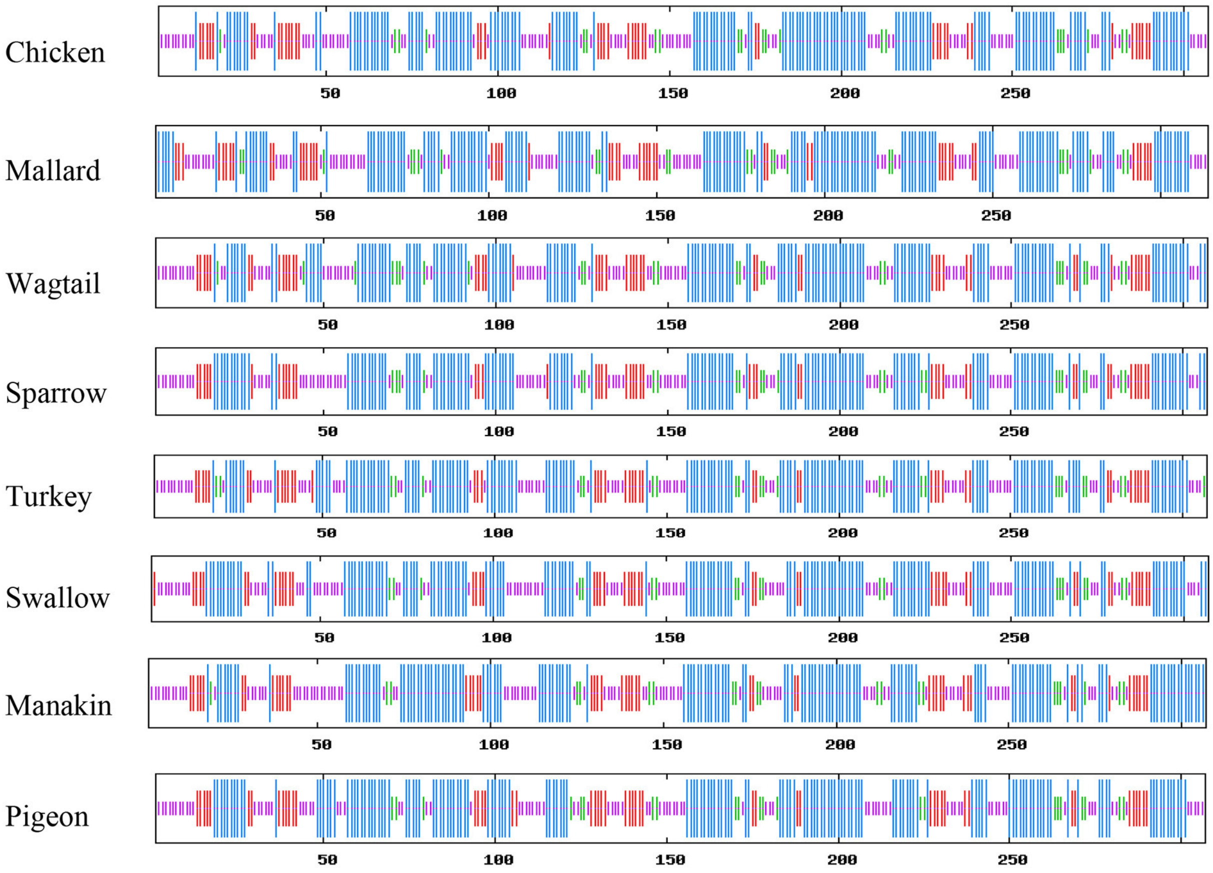


FIGURE 5
Comparative view of the secondary structure of *avUCPs* in eight avian species (Blue, Alpha helix; Purple, Random coil; Red, Extended-strand; Green, Beta turn).

protein was represented as the reliability of GO prediction. The GO trees of *avUCP* are depicted in Figure 7. Consequently, four biological processes including "mitochondrial transmembrane transport," "proton transmembrane transport," "adaptive thermogenesis," and "response to cold" were predicted with 57% reliability. Also, three cellular components including "mitochondrion," "mitochondrial inner membrane," and "integral component of the membrane," along with one molecular function of "oxidative phosphorylation uncoupler activity" were also anticipated with 57% confidence.

3.5. Interactive network prediction and gene set enrichment analysis

Gene network was predicted in *Cytoscape* software (3.9.1) using the embedded *STRING* app, and clustering was performed by the K-means method (72). A total of 49 published *avUCP* related-genes were used as input for network prediction and resulted in highly orchestrated interactions among genes. This network is divided into two clusters with the highest confidence: cluster one containing 21 genes involved in response to stress, and cluster two containing 28 genes involved in lipid metabolism and proton buffering system. The predicted network is shown in Figure 8. GO analysis on these networked genes highlighted several biological processes including the fatty acid metabolic process (GO:0006631), response to chemical (GO:0042221), cellular response to chemical stimulus (GO:0070887), oxidation-reduction process (GO: 0055114), fatty acid beta-oxidation (GO: 0006635), and regulation of fatty acid metabolic process (GO:0019217) as being enriched (FDR < 0.01). Furthermore, three cellular components including "mitochondrion," "TOR complex1," and "mitochondrial membrane," were enriched (FDR $<$ 0.01). The PPAR signaling pathway, adipocytokine signaling, FoXO signaling, mTOR signaling pathway, insulin signaling pathway, MAPK signaling pathway, fatty acid degradation, fatty acid metabolism, GnRH signaling pathway, ErbB signaling pathway, and oxidative phosphorylation were also indicated as enriched KEGG pathways for the *avUCP* gene network (FDR < 0.05).

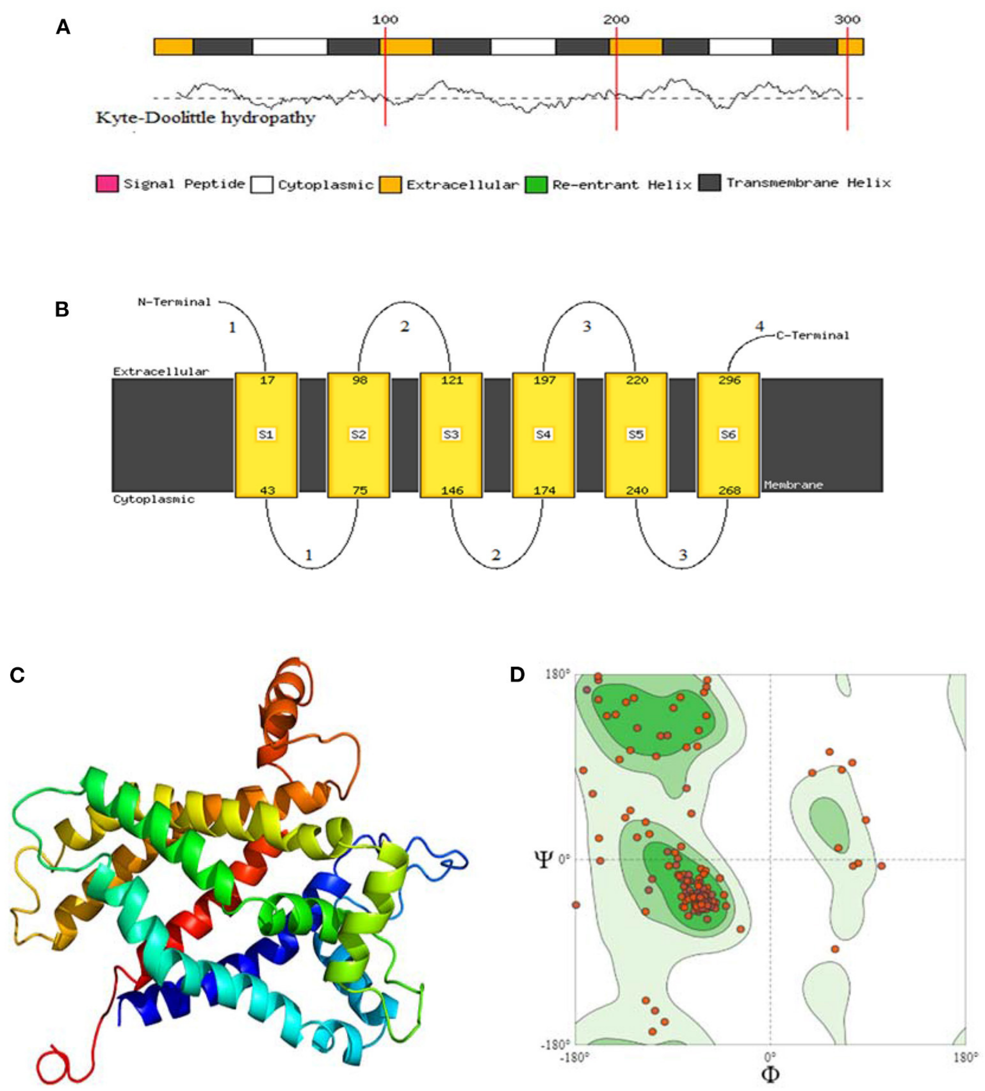


FIGURE 6
(A) Membrane helix prediction with support vector machines with Kyte-Doolittle hydropathy plot. **(B)** Schematic figure of predicted extracellular, cytoplasmic, and transmembrane helices. **(C)** Top 3D model of avian *UCP* [Model dimensions (Å): X: 53.347, Y: 74.255, Z: 50.565, Image colored by the rainbow N → C terminus]. **(D)** Ramachandran plot with 88.46% favored. A dihedral angle of a protein is the internal angle of polypeptide backbone at which two adjacent planes meet. The conformation of the backbone can be described by two dihedral angles per residue, because the backbone residing between two juxtaposing Cα atoms are all in a single plane. These angles are called ϕ (phi) which involves the backbone atoms C-N-Cα-C, and ψ (psi) which involves the backbone atoms N-Cα-C-N.

3.6. Pathway crosstalk

After a query of enriched pathways, we used XtalkDB (75) to predict which pairs of signaling pathways may crosstalk with each other. The six enriched pathways were predicted to be involved in a network of crosstalk which is depicted in Figure 9. In detail, adipocytokine signaling was revealed to have an activation effect on MAPK, insulin, and ErbB signaling pathways, and both activation and inhibition effects on the mTOR signaling pathway.

The mTOR signaling pathway is predicted to act as an inhibitor for both MAPK and insulin signaling pathways. Among these, the MAPK showed only an activation effect on the mTOR pathway. Moreover, the insulin signaling pathway is anticipated to induce adipocytokine, mTOR, and ErbB signaling pathways, alongside inhibited GnRH signaling pathway. The GnRH signaling pathway is shown to be involved in activation of adipocytokine and MAPK signaling pathways and both negative and positive crosstalk with ErbB signaling pathway. In addition, the ErbB signaling pathway can activate and silence MAPK and insulin signaling pathways, respectively.

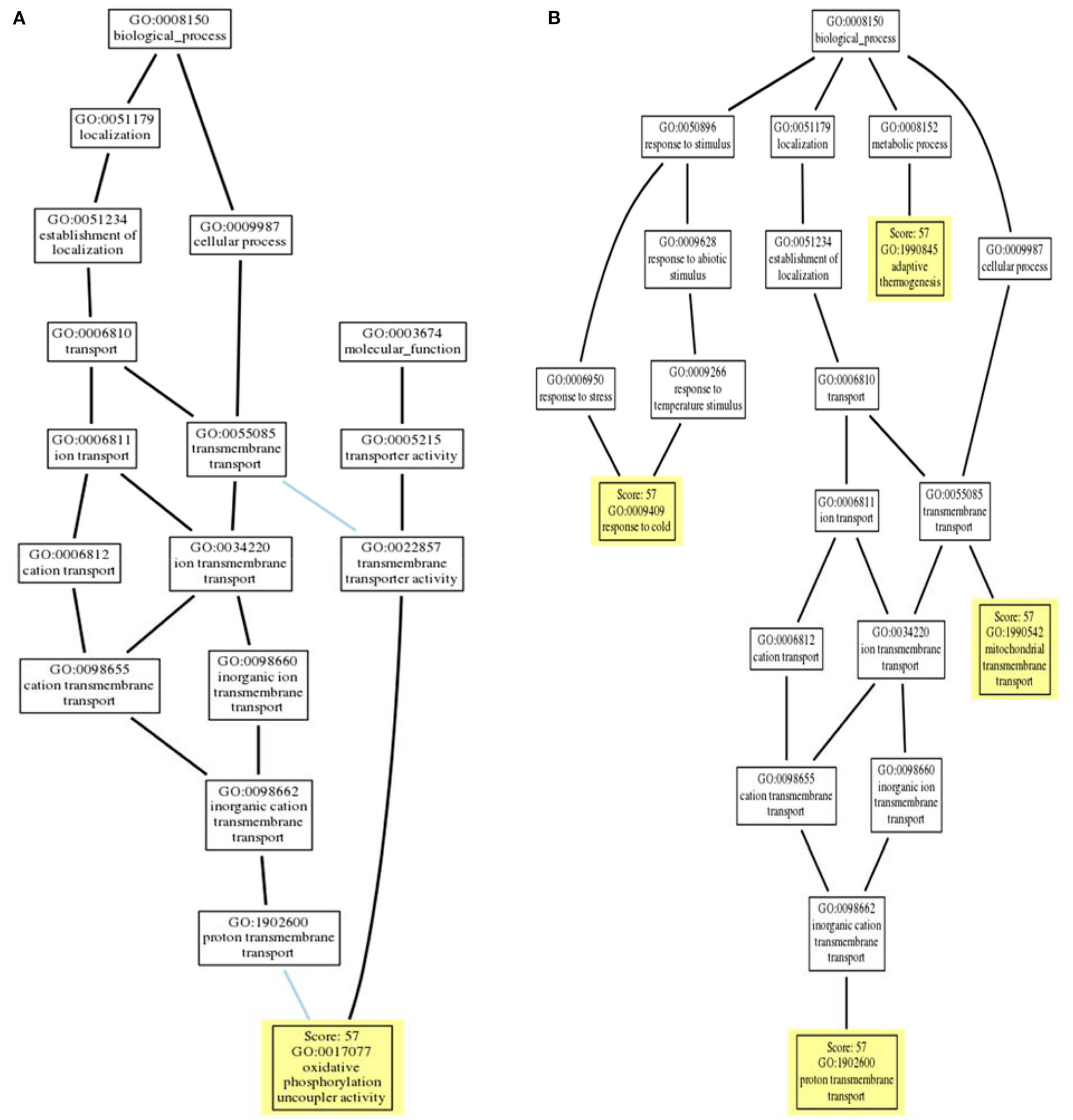

A
GO:0008150 biological_process
GO:0051179 localization
GO:0051234 establishment of localization
GO:0009987 cellular process
GO:0006810 transport
GO:0003674 molecular_function
GO:0006811 ion transport
GO:0055085 transmembrane transport
GO:0005215 transporter activity
GO:0006812 cation transport
GO:0034220 ion transmembrane transport
GO:0022857 transmembrane transporter activity
GO:0098655 cation transmembrane transport
GO:0098660 inorganic ion transmembrane transport
GO:0098662 inorganic cation transmembrane transport
GO:1902600 proton transmembrane transport
Score: 57 GO:0017077 oxidative phosphorylation uncoupler activity
B
GO:0008150 biological_process
GO:0050896 response to stimulus
GO:0051179 localization
GO:0008152 metabolic process
GO:0009628 response to abiotic stimulus
GO:0051234 establishment of localization
Score: 57 GO:1990845 adaptive thermogenesis
GO:0009987 cellular process
GO:0006950 response to stress
GO:0009266 response to temperature stimulus
GO:0006810 transport
Score: 57 GO:0009409 response to cold
GO:0006811 ion transport
GO:0055085 transmembrane transport
GO:0006812 cation transport
GO:0034220 ion transmembrane transport
Score: 57 GO:1990542 mitochondrial transmembrane transport
GO:0098655 cation transmembrane transport
GO:0098660 inorganic ion transmembrane transport
GO:0098662 inorganic cation transmembrane transport
Score: 57 GO:1902600 proton transmembrane transport

FIGURE 7
(A) Molecular function ontology tree and **(B)** biological process ontology tree (locations of PREDICTED terms are highlighted in yellow with respect to inferred terms).

4. Discussion

4.1. Coding sequence analysis

4.1.1. Codon usage analysis

In the investigation of codon usage, the RSCU value is a ratio between the occurrence frequency of a certain codon and the expected usage frequency for codons (78). Codons encoding amino acids of *avUCP* with RSCU values higher than 1.0 represent positive codon usage bias and codons with RSCU values lower than 1 display negative codon usage bias. Moreover, codons with RSCU = 0 display unfavorable codons. According to the results, the *avUCP* gene obviously prefers codons with "C" and "G" in the third position over the other bases. It can

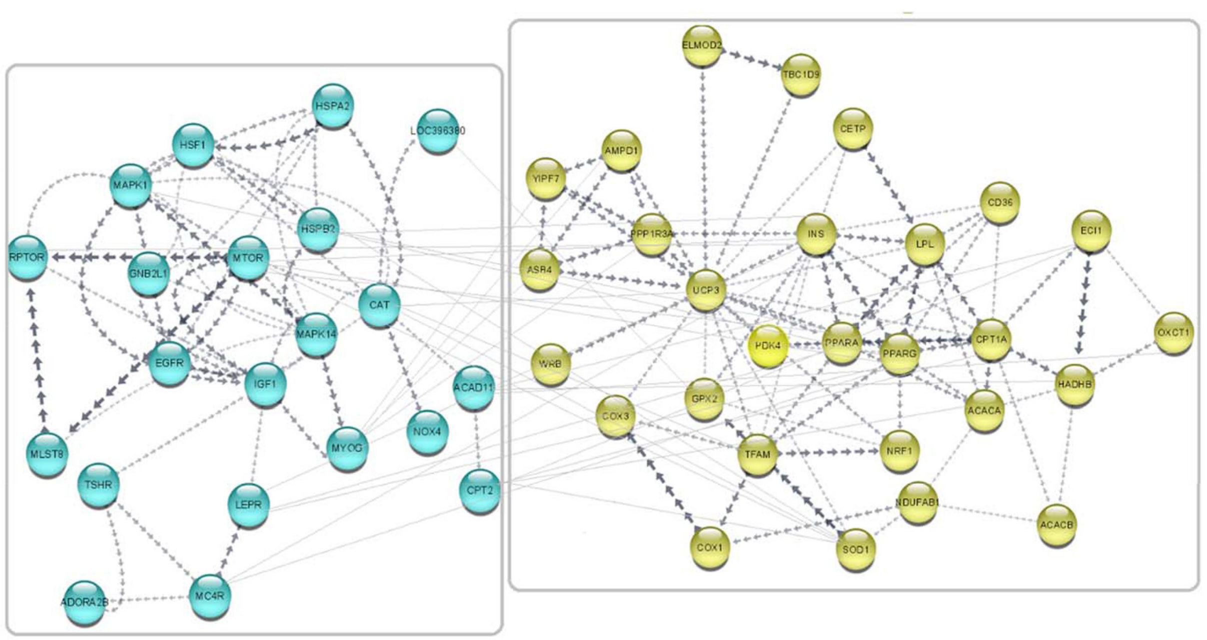


FIGURE 8
Predicted network for the regulatory and collaborative genes with *avUCP*. Separate boxes show clusters. Blue nodes represent genes involved in response to stress (cold and free radicals) (cluster 1). Yellow nodes represent genes involved in lipid metabolism and proton buffering system (cluster 2).

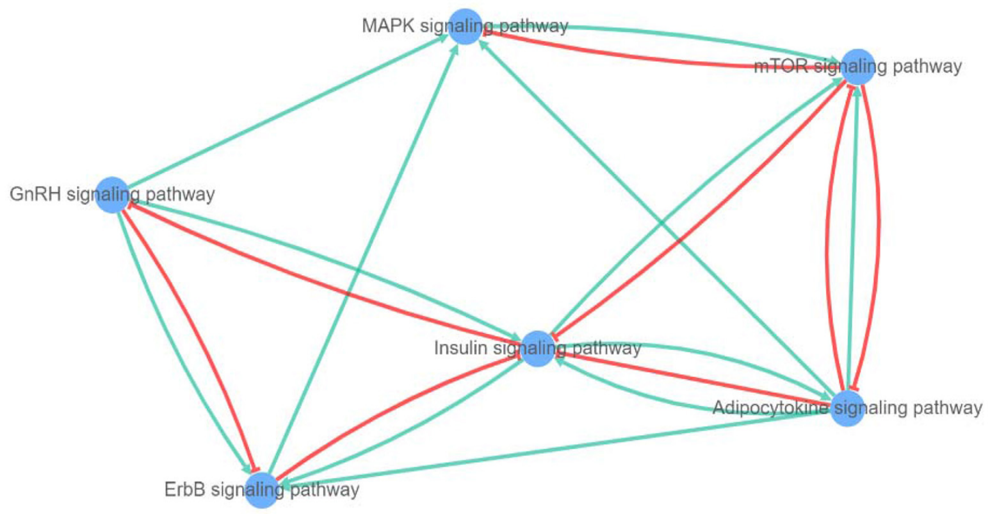


FIGURE 9
Network of crosstalk among enriched pathways (blue lines show activation and red lines represent inhibition effect).

be concluded that almost all highly preferred codons, except the CAU codon (only in manakin), which encodes for histidine, end with a "C" or a "G."

However, the mechanisms of inducing codon biases remained an open question. It can be attributed to the expression level of genes, selective pressure, evolutionary trend, phylogenetic relations of organisms, and genetic drift (79–81).

Moreover, the CAI values of 1 and 0 refer to the species in which only the most frequent codons are used, and species using the least frequent codons. Although manakin had the lowest GC ratio among the studied birds, it revealed the highest CAI, indicating that it uses 71.2% of the frequently used codons. Manakin uses most of the synonymous codons (53/61 codons) to encode *avUCP* so that it has the lowest number of unused codons (8). Interestingly, manakin is the only species among the selected birds which breeds in tropical forests. Therefore, the observed differences in manakin may be a result of adaptation to tropical conditions. Furthermore, the use of a comparatively wide range of synonymous codons alongside the use of frequently-used codons can be regarded as an evolutionary variation to achieve efficient translation in relatively important functional genes.

4.2. Protein sequence analysis

In protein sequence and physio-chemical analysis, a slight variation in amino acids and physio-chemical parameters among the eight studied *avUCP* proteins was detected. Tyrosine was the only amino acid showing a constant level in all avian species. The range of isoelectric points of *UCP*s implies that *UCP*s can be membrane proteins. As the pH of the intermembrane space and the mitochondrial matrix is about 7.0 and 8.0, respectively, thus *UCP*s carry an electrical charge in that region. Additionally, since aliphatic side chains like alanine, leucine, and valine determine the aliphatic index (AI) of a protein, *avUCP,* which shows a high content of these amino acids, could be considered thermostable. The highest AI of 100.42 was observed in manakin *avUCP* which may illustrate the importance of *avUCP* stability specifically in manakin.

The variation of atomic sulfur count from 14 to 20 among the eight *avUCP* proteins illustrates another significant difference among avian species. Sulfur can be found in cysteine and methionine amino acids. Nine cysteine residues were observed in *avUCP* in sparrow but the protein sequence of the others contained eight cysteine residues. Furthermore, the hydrophobic methionine was variable among *avUCP*s, which could be the second source of variability in atomic sulfur count among *avUCP* sequences in the studied species. Sulfur-containing amino acids are responsible for stronger connections than aliphatic and aromatic amino acids. Thus, they can provide a more sustainable 3D structure representing functional specificity in membrane proteins by creating a disulfide bond (77, 82). The percentage of hydrophobic and hydrophilic amino acids in the studied sequences were very close to each other indicating that these chemical characteristics play important roles in the encoded *avUCP* protein. The higher potential of hydrophobicity according to GRAVY indices above 0 is another appropriate state for transmembrane proteins. If a protein plays an important functional role, the state of its hydrophobicity and hydrophilicity will remain stable as much as possible, for the conservation of its function. Also, the stability of positively and negatively charged amino acids among all studied *avUCP*s may imply the effectiveness of charged regions of this protein which needs to be conserved among birds.

4.2.1. Entropy analysis

The positions of aligned sequences in the entropy plot can correlate with the structural and chemical characteristics of certain amino acids and their influence on the function of *avUCP*s. Therefore, regions that contain residue positions with low entropy are more likely to be involved in the functional sites of *avUCP* (83, 84). Moreover, the absence of any region with more than 10 consecutive amino acids with an average entropy of >1 can be concluded as evidence of conservation in *avUCPs*.

4.3. Protein structure prediction

Because of the high frequency of alanine, leucine, valine, and glycine, and the low frequency of histidine and tryptophan in all *avUCP*s, the level of alpha-helical structure was determined higher than in other secondary structures. In agreement with our result, it is already known that transmembrane regions of proteins contain a high level of alpha-helices devoid of polar amino acids, while extracellular and cytoplasmic regions of the protein are usually enriched with polar amino acids like tyrosine and tryptophan (77).

The Ramachandran parameters of predicted protein can display the statistical distribution of the combinations of torsional *Phi* and *Psi* angles in the *avUCP* protein structure (85). Moreover, rotamers in protein structure imply conformational isomers of amino acid residues in the sidechain of *avUCP*, therefore, rotamer outliers display conformations that drop outside the reference (86). Also, C-beta deviation can reflect misfit conformation and inconsistency between sidechain and protein backbone that can be used for structure validation (87). In the current structural modeling for *avUCP*, the two C-beta deviations have resulted from valine and proline amino acids in positions 56 and 50, respectively, which were predicted to be in the cytoplasmic region of *avUCP*.

Hence, in addition to regulatory hormones and elements, different innate parameters can affect gene expression patterns of any pleiotropic genes like *avUCP,* with those parameters including, codon usage, GC content, CpG dinucleotide content, splicing sites, CpG islands, mRNA secondary structure, coding sequences (CDS), ribosomal binding sites, stimulators, the expression of other genes, along with environmental conditions (78, 88). For example, previous research has revealed cysteine residues of *UCP*s can be glutathionylated (89–92). They suggest that reactive oxygen species and glutathionylation can regulate non-phosphorylating respiration. Mailloux et al. (93) have identified Cys^{25} and Cys^{259} as the probable glutathionylation sites on *UCPs* (93). Interestingly, in the current study Cys^{25} and Cys^{257} were determined as conserved sites and predicted to be located in transmembrane and cytoplasmic regions of avUCP, respectively.

4.4. Sequence-based gene ontology prediction

The sequence-based predicted biological processes of "mitochondrial transmembrane transport," "proton transmembrane transport," "adaptive thermogenesis," and "response to cold" for *avUCP* are congruent with results from previous studies. In this regard, some studies support the involvement of *avUCP* in avian energy expenditure and adaptive thermogenesis (1, 15–18). Additionally, it should be mentioned that cold acclimation can not only induce fatty acid-mediated uncoupling of oxidative phosphorylation processes but also increases the rate of ADP and Pi concentrations, along with ATP synthesis in the mitochondria of chicken skeletal muscle, which seems to be a counterproductive occurrence in response to cold stress condition in birds (34). Moreover, Ueda et al. reported a correlation between uncoupling and both exogenous and endogenous fatty acids in the mitochondria of chicken skeletal muscle during cold temperatures (94). Another study conducted on king penguins showed that superoxide activates the proton transport of mitochondria and GDP inhibits the transport of the superoxide-activated-proton, demonstrating that *avUCP* mediates mitochondrial proton transport but plays no role in the basal proton leak (6). Because thermogenic hormones have an induction effect on *avUCP* expression, the involvement of avUCP in avian thermogenesis can be concluded (16).

4.5. Interactive network prediction and gene-based enrichment analysis

Through the network of gene-gene interactions, centralized by *avUCP* and using previously recognized *avUCP*- related genes (5, 34, 42, 56–71, 94), we have found two major clusters of genes pointing to the overall functionality of response to stress and lipid metabolism/proton buffering. Our well-categorized findings are in agreement with the outputs of other studies. Previously conducted studies revealed the predominant presence of *avUCP* protein in skeletal muscles (pectoral, glycolytic fibers) (90) alongside recognizing the alteration in expression pattern during different physiological states [cold stress (10, 12, 34, 58, 90), heat stress (12, 29, 92, 95, 96), transfer stress (97), high fat diet, and fat (16, 38, 66, 95)], that clearly reflect the involvement of *avUCP* in fatty acid β-oxidation and cell metabolism. Moreover, one study reported that a high concentration of chemical stimulus like ammonia can be effective in the expression of 12 energy metabolism-related genes (*avUCP, HK1*, *HK2*, *PK*, *PFK*, *PDHX*, *CS*, *LDHA*, *LDHB*, *SDHA*, *SDHB*, and *AMPK*), in chicken liver. Otherwise, it was reported that ammonia gas resulted in mitochondrial damage, ATPase reduction, and ultimately reduction of energy release in the chicken liver (98).

4.6. Pathway crosstalk

Finally, among 11 enriched pathways, interaction of five signaling pathways including MAPK, adipocytokine, mTOR, insulin, ErbB, and GnRH was predicted, indicating a possible combination of positive and negative feedback among pathways to regulate *avUCP* functions. In general, biological pathway crosstalk refers to the different feedback in seemingly distinct pathways (99). Consequently, it seems that maintaining a delicate balance of *avUCP* functions such as lipid metabolism, thermogenesis, response to cold, and response to ROS can occur by crosstalk between involved pathways. Moreover, when a single gene is considered in depth, within a network of genes, all potential regulatory interferences will emerge. Additionally, it is also known that the interaction between pathways can regulate specific gene expression (100, 101).

A panoply of changes in the primary sequence of *avUCPs* can potentially be involved in changes to protein function and expression through alteration of the final structure of the *avUCP* molecule. Accumulation of findings represents *avUCP* as an essential gene for whole-body energy balance, adaptive thermogenesis, and antioxidant defense in birds. This study contributes to a better understanding of *avUCP* characterization, function, and critical signaling pathways for evaluating how it is regulated in avian species exposed to different conditions. Additionally, the present study provides putative functions for *avUCP*s, and indicates some of the genes, pathways, and mechanisms that are involved in fine-tuning mitochondrial oxidative phosphorylation.

In conclusion, we have compared the sequence, structure and physio-chemical properties of *avUCP* in 8 bird species and determine the functional pathways and networks in which *avUCP* is involved. Oxidative stress in birds is known as one of the most energy-demanding events influencing energy expenditure, the balance of detoxification of free radicals, and oxidative phosphorylation, so *avUCP* could be viewed as a significant marker for developing heat-stress-resistant breeds in future genomic selection programmes.

Data availability statement

The data presented in the study (nucleotide coding sequences and amino acid sequences of avian uncoupling proteins in FASTA format) are deposited at https://figshare.com/ with the accession number https://figshare.com/search?q=10.6084%2Fm9.figshare.20518086 and are also presented in Table 1.

Author contributions

PD and MG-Z conceived the overall design, undertook the project management, contributed to the interpretation of results, and critically revised the manuscript. PD carried out the analyses and drafted the manuscript. MD, MR, SK, and EE performed the review part of this research, collecting a set of genes used for gene network analysis and provided support in the analyses. JS revised the entire manuscript, contributed to the interpretation of results, and critically reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded in part by Biotechnology and Biological Sciences Research Council (BBSRC) grant number BBS/E/D/10002071.

Acknowledgments

The authors are grateful for the insightful comments offered by the peer reviewers at the Journal of Frontiers in Veterinary Science. The authors would also wish to acknowledge Dr. Helene Hayes and Jack Dekkers for their support and encouragement. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Raimbault S, Dridi S, Denjean F, Lachuer J, Couplan E, Bouillaud F, et al. An uncoupling protein homologue putatively involved in facultative muscle thermogenesis in birds. *Biochem J.* (2001) 353:441–5. doi: 10.1042/bj3530441

2. Darzi Niarami M, Masoudi AA, Vaez Torshizi R, Davoodi P. A novel mutation in the promoter region of avian uncoupling protein3 associated with feed efficiency and body composition traits in broiler chicken. *J Worlds Poult Res.* (2020) 10:87–95. doi: 10.36380/jwpr.2020.12

3. Beltran DF, Shultz AJ, Parra JL. Speciation rates are positively correlated with the rate of plumage color evolution in hummingbirds. *Evolution.* (2021) 75:1665–80. doi: 10.1111/evo.14277

4. Vianna C, Hagen T, Zhang C-Y, Bachman E, Boss O, Gereben B, et al. Cloning and functional characterization of an uncoupling protein homolog in hummingbirds. *Physiol Genomics.* (2001) 5:137–45. doi: 10.1152/physiolgenomics.2001.5.3.137

5. Evock-Clover CM, Poch SM, Richards MP, Ashwell CM, Mcmurtry JP. Expression of an uncoupling protein gene homolog in chickens. *Comp Biochem Physiol B Biochem.* (2002) 133:345–58. doi: 10.1016/S1095-6433(02)00113-7

6. Talbot DA, Hanuise N, Rey B, Rouanet JL, Duchamp C, Brand MD. Superoxide activates a GDP-sensitive proton conductance in skeletal muscle mitochondria from king penguin (*Aptenodytes patagonicus*). *Biochem Biophys Res Commun.* (2003) 312:983–8. doi: 10.1016/j.bbrc.2003.11.022

7. Dridi S, Onagbesan O, Swennen Q, Buyse J, Decuypere E, Taouis M. Gene expression, tissue distribution and potential physiological role of uncoupling protein in avian species. *Comp Biochem Physiol A Mol Integr Physiol.* (2004) 139:273–83. doi: 10.1016/j.cbpb.2004.09.010

8. Duval E. Age-based plumage changes in the Lance-tailed Manakin: a two-year delay in plumage maturation. *Condor.* (2009) 107:915–20. doi: 10.1093/condor/107.4.915

9. Garratt M, Brooks RC. Oxidative stress and condition-dependent sexual signals: more than just seeing red. *Proc Biol Sci.* (2012) 279:3121–30. doi: 10.1098/rspb.2012.0568

10. Wollenberg Valero KC, Pathak R, Prajapati I, Bankston S, Thompson A, Usher J, et al. A candidate multimodal functional genetic network for thermal adaptation. *Peer J.* (2014) 2:e578. doi: 10.7717/peerj.578

11. Davoodi P, Ehsani A, Vaez Torshizi R, Masoudi AA. New insights into genetics underlying of plumage color. *Anim Genet.* (2021) 53:80–93. doi: 10.1111/age.13156

12. Mota-Rojas D, Titto CG, De Mira Geraldo A, Martínez-Burnes J, Gómez J, Hernández-Ávalos I, et al. Efficacy and function of feathers, hair, and glabrous skin in the thermoregulation strategies of domestic animals. *Animals.* (2021) 11:3472. doi: 10.3390/ani11123472

13. Hohtola E, Henderson RP, Rashotte ME. Shivering thermogenesis in the pigeon: the effects of activity, diurnal factors, and feeding state. *Am J Physiol Regul Integr Comp Physiol AM.* (1998) 275:R1553–62. doi: 10.1152/ajpregu.1998.275.5.R1553

14. Bicudo JE, Vianna CR, Chaui-Berlinck JG. Thermogenesis in birds. *Biosci Rep.* (2001) 21:181–8. doi: 10.1023/A:1013648208428

15. Taouis M, De Basilio V, Mignon-Grasteau S, Crochet S, Bouchot C, Bigot K, et al. Early-age thermal conditioning reduces uncoupling protein messenger RNA expression in pectoral muscle of broiler chicks at seven days of age. *Poult Sci.* (2002) 81:1640–3. doi: 10.1093/ps/81.11.1640

16. Collin A, Taouis M, Buyse J, Ifuta NB, Darras VM, Van As P, et al. Thyroid status, but not insulin status, affects expression of avian uncoupling protein mRNA in chicken. *Am J Physiol Endocrinol.* (2003) 284:E771–8. doi: 10.1152/ajpendo.00478.2002

17. Collin A, Cassy S, Buyse J, Decuypere E, Damon M. Potential involvement of mammalian and avian uncoupling proteins in the thermogenic effect of thyroid hormones. *Domest Anim Endocrinol.* (2005) 29:78–87. doi: 10.1016/j.domaniend.2005.02.007

18. Mujahid A. Acute cold-induced thermogenesis in neonatal chicks (*Gallus gallus*). *Comp Biochem Physiol Part A Mol Integr Physiol.* (2010) 156:34–41. doi: 10.1016/j.cbpa.2009.12.004

19. Swennen Q, Delezie E, Collin A, Decuypere E, Buyse J. Further investigations on the role of diet-induced thermogenesis in the regulation of feed intake in chickens: comparison of age-matched broiler versus layer cockerels. *Poult Sci.* (2007) 86:895–903. doi: 10.1093/ps/86.5.895

20. Teulier L, Rouanet JL, Letexier D, Romestaing C, Belouze M, Rey B, et al. Cold-acclimation-induced non-shivering thermogenesis in birds is associated with upregulation of avian UCP but not with innate uncoupling or altered ATP efficiency. *J Exp Biol.* (2010) 213:2476–82. doi: 10.1242/jeb.043489

21. Rey B, Halsey LG, Dolmazon V, Rouanet J-L, Roussel D, Handrich Y, et al. Long-term fasting decreases mitochondrial avian UCP-mediated oxygen consumption in hypometabolic king penguins. *Am J Physiol Regul Integr Comp Physiol.* (2008) 295:92–100. doi: 10.1152/ajpregu.00271.2007

22. Jin S, Yang L, He T, Fan X, Wang Y, Ge K, et al. Polymorphisms in the uncoupling protein 3 gene and their associations with feed efficiency in chickens. *Asian Australas J Anim Sci.* (2018) 31:1401–6. doi: 10.5713/ajas.18.0217

23. Abe T, Mujahid A, Sato K, Akiba Y, Toyomizu M. Possible role of avian uncoupling protein in down-regulating mitochondrial superoxide production in skeletal muscle of fasted chickens. *FEBS Lett.* (2006) 580:4815–22. doi: 10.1016/j.febslet.2006.07.070

24. Criscuolo F, Mozo J, Hurtaud C, Nübel T, Bouillaud F. UCP2, UCP3, avUCP, what do they do when proton transport is not stimulated*? Possible relevance to pyruvate and glutamine metabolism. Biochim Biophys Acta.* (2006) 1757:1284–91. doi: 10.1016/j.bbabio.2006.06.002

25. Mujahid A, Sato K, Akiba Y, Toyomizu M. Acute heat stress stimulates mitochondrial superoxide production in broiler skeletal muscle, possibly via downregulation of uncoupling protein content. *Poult Sci.* (2006) 85:1259–65. doi: 10.1093/ps/85.7.1259

26. Mujahid A, Akiba Y, Toyomizu M. Acute heat stress induces oxidative stress and decreases adaptation in young white leghorn cockerels by downregulation of avian uncoupling protein. *Poult Sci.* (2007) 86:364–71. doi: 10.1093/ps/86.2.364

27. Del Vesco AP, Gasparino E. Production of reactive oxygen species, gene expression, and enzymatic activity in quail subjected to acute heat stress. *J Anim Sci.* (2013) 91:582–7. doi: 10.2527/jas.2012-5498

28. Kikusato M, Toyomizu M. Crucial role of membrane potential in heat stress-induced overproduction of reactive oxygen species in avian skeletal muscle mitochondria. *PLoS ONE.* (2013) 8:e64412. doi: 10.1371/journal.pone.0064412

29. Lu Z, He X, Ma B, Zhang L, Li J, Jiang Y, et al. Chronic heat stress impairs the quality of breast-muscle meat in broilers by affecting redox status and energy-substance metabolism. *J Agric Food Chem.* (2017) 65:11251–8. doi: 10.1021/acs.jafc.7b04428

30. Xu L, Wu SG, Zhang HJ, Zhang L, Yue HY, Ji F, et al. Comparison of lipid oxidation, messenger ribonucleic acid levels of avian uncoupling protein, avian adenine nucleotide translocator, and avian peroxisome proliferator-activated receptor-γ coactivator-1α in skeletal muscles from electrical- and gas-stunned broilers. *Poult Sci.* (2011) 90:2069–75. doi: 10.3382/ps.2011-01348

31. Liu J-S, Chen Y-Q, Li M. Thyroid hormones increase liver and muscle thermogenic capacity in the little buntings (*Emberiza pusilla*). *J Therm Biol.* (2006) 31:386–93. doi: 10.1016/j.jtherbio.2006.01.002

32. Ferver A, Dridi S. Regulation of avian uncoupling protein (av-UCP), expression by cytokines and hormonal signals in quail myoblast cells. *Comp Biochem Physiol Part A Mol Integr Physiol.* (2020) 248:110747. doi: 10.1016/j.cbpa.2020.110747

33. Bal NC, Periasamy M. Uncoupling of sarcoendoplasmic reticulum calcium ATPase pump activity by sarcolipin as the basis for muscle non-shivering thermogenesis. *Philos Trans R Soc Lond, B, Biol Sci.* (2020) 375:20190135. doi: 10.1098/rstb.2019.0135

34. Toyomizu M, Ueda M, Sato S, Seki Y, Sato K, Akiba Y. Cold-induced mitochondrial uncoupling and expression of chicken UCP and ANT mRNA in chicken skeletal muscle. *FEBS Lett.* (2002) 529:313–8. doi: 10.1016/S0014-5793(02)03395-1

35. Walter I, Seebacher F. Endothermy in birds: underlying molecular mechanisms. *J Exp Biol.* (2009) 212:2328–36. doi: 10.1242/jeb.029009

36. Collin A, Buyse J, Van As P, Darras VM, Malheiros RD, Moraes VM, et al. Cold-induced enhancement of avian uncoupling protein expression, heat production, and triiodothyronine concentrations in broiler chicks. *Gen Comp Endocrinol.* (2003) 130:70–7. doi: 10.1016/S0016-6480(02)00571-3

37. Zheng S, Zhao J, Xing H, Xu S. Oxidative stress, inflammation, and glycometabolism disorder-induced erythrocyte hemolysis in selenium-deficient exudative diathesis broilers. *J Cell Physiol*. (2019). doi: 10.1002/jcp.28298

38. Collin A, Swennen Q, Skiba-Cassy S, Buyse J, Chartrin P, Le Bihan-Duval E, et al. Regulation of fatty acid oxidation in chicken (*Gallus gallus*): interactions between genotype and diet composition. *Comp Biochem Physiol B.* (2009) 153:171–7. doi: 10.1016/j.cbpb.2009.02.012

39. Joubert R, Métayer Coustard S, Swennen Q, Sibut V, Crochet S, Cailleau-Audouin E, et al. The beta-adrenergic system is involved in the regulation of the expression of avian uncoupling protein in the chicken. *Domest Anim Endocrinol.* (2010) 38:115–25. doi: 10.1016/j.domaniend.2009.08.002

40. Do PH, Tran PV, Bahry MA, Yang H, Han G, Tsuchiya A, et al. Oral administration of a medium containing both D-aspartate-producing live bacteria and D-aspartate reduces rectal temperature in chicks. *Br Poult Sci.* (2017) 58:569–77. doi: 10.1080/00071668.2017.1335858

41. Kikusato M, Muroi H, Uwabe Y, Furukawa K, Toyomizu M. Oleuropein induces mitochondrial biogenesis and decreases reactive oxygen species generation in cultured avian muscle cells, possibly via an up-regulation of peroxisome proliferator-activated receptor γ coactivator-1α. *Anim Sci J.* (2016) 87:1371–8. doi: 10.1111/asj.12559

42. Muroi H, Hori K, Tokutake Y, Hakamata Y, Kawabata F, Toyomizu M, et al. Oleuropein suppresses mitochondrial reactive oxygen species generation possibly via an activation of transient receptor potential V1 and sirtuin-1 in cultured chicken muscle cells. *Anim Sci J.* (2022) 93:e13677. doi: 10.1111/asj.13677

43. Lassiter K, Greene E, Piekarski A, Faulkner OB, Hargis BM, Bottje W, et al. Orexin system is expressed in avian muscle cells and regulates mitochondrial dynamics. *Am J Physiol Regul Integr Comp Physiol.* (2015) 308:173–87. doi: 10.1152/ajpregu.00394.2014

44. Criscuolo F, Gonzalez-Barroso MDM, Maho YL, Ricquier D, Bouillaud F. Avian uncoupling protein expressed in yeast mitochondria prevents endogenous free radical damage. *Proc Royal Soc B.* (2005) 272:803–10. doi: 10.1098/rspb.2004.3044

45. Yalin E, Hurtaud C, Ricquier D, Bouillaud F, Hughes J, Criscuolo F. Avian UCP: the killjoy in the evolution of the mitochondrial uncoupling proteins. *J Mol Evol.* (2007) 65:392–402. doi: 10.1007/s00239-007-9020-1

46. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol.* (2021) 38:3022–7. doi: 10.1093/molbev/msab120

47. Puigbò P, Bravo IG, Garcia-Vallvé S. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinform*. (2008) 9:65. doi: 10.1186/1471-2105-9-65

48. Qiagen. *CLC Genomics Workbench 20.0 (QIAGEN).* (2016). Available online at: www.qiagenbioinformatics.com (accessed June 14, 2022).

49. Alzohairy A. BioEdit: an important software for molecular biology. *GERF Bull Biosci.* (2011) 2:60–1.

50. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics.* (2014) 15:7. doi: 10.1186/1471-2105-15-7

51. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci.* (1995) 11:681–4. doi: 10.1093/bioinformatics/11.6.681

52. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* (2015) 10:845–58. doi: 10.1038/nprot.2015.053

53. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* (2018) 46:296–303. doi: 10.1093/nar/gky427

54. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. Predict protein – predicting protein structure and function for 29 years. *BioRxiv*. (2021). doi: 10.1093/nar/gkab354

55. Davoodi P, Ehsani A. *In*-silico investigation of genomic regions related to ascites and identifying their pathways in broilers. *World's Poult Sci J.* (2019) 75:193–206. doi: 10.1017/S0043933919000035

56. Fromme T, Reichwald K, Platzer M, Li X-S, Klingenspor M. Chicken ovalbumin upstream promoter transcription factor II regulates uncoupling protein 3 gene transcription in Phodopus sungorus. *BMC Mol Biol.* (2007) 8:1. doi: 10.1186/1471-2199-8-1

57. Dridi S, Temim S, Derouet M, Tesseraud S, Taouis M. Acute cold- and chronic heat-exposure upregulate hepatic leptin and muscle uncoupling protein (UCP), gene expression in broiler chickens. *J Exp Zool A Ecol Genet Physiol.* (2008) 309:381–8. doi: 10.1002/jez.461

58. Wang JT, Zhang XJ, Xu SW. Effects of cold stress on energy metabolism in the chicken. *J Appl Physiol.* (2009) 25:172–6.

59. Cai Y, Song Z, Wang X, Jiao H, Lin H. Dexamethasone-induced hepatic lipogenesis is insulin dependent in chickens (*Gallus gallus* domesticus). *Stress.* (2011) 14:273–81. doi: 10.3109/10253890.2010.543444

60. Sibut V, Hennequet-Antier C, Le Bihan-Duval E, Marthey S, Duclos MJ, Berri C. Identification of differentially expressed genes in chickens differing in muscle glycogen content and meat quality. *BMC Genom.* (2011) 12:112. doi: 10.1186/1471-2164-12-112

61. Cui HX, Liu RR, Zhao GP, Zheng MQ, Chen JL, Wen J. Identification of differentially expressed genes and pathways for intramuscular fat deposition in pectoralis major tissues of fast-and slow-growing chickens. *BMC Genom.* (2012) 13:213. doi: 10.1186/1471-2164-13-213

62. Ji B, Ernest B, Gooding JR, Das S, Saxton AM, Simon J, et al. Transcriptomic and metabolomic profiling of chicken adipose tissue

in response to insulin neutralization and fasting. *BMC Genom.* (2012) 13:441. doi: 10.1186/1471-2164-13-441

63. Li Q, Xu Z, Liu L, Yu H, Rong H, Tao L, et al. Effects of breeds and dietary protein levels on the growth performance, energy expenditure and expression of avUCP mRNA in chickens. *Mol Biol Rep*. (2013) 40:2769–79. doi: 10.1007/s11033-012-2030-0

64. Saneyasu T, Shiragaki M, Nakanishi K, Kamisoyama H, Honda K. Effects of short term fasting on the expression of genes involved in lipid metabolism in chicks. *Comp Biochem Physiol B, Biochem Mol Biol.* (2013) 165:114–8. doi: 10.1016/j.cbpb.2013.03.005

65. Hicks JA, Porter TE, Liu HC. Identification of microRNAs controlling hepatic mRNA levels for metabolic genes during the metabolic transition from embryonic to posthatch development in the chicken. *BMC Genom.* (2017) 18:687. doi: 10.1186/s12864-017-4096-5

66. Desert C, Baéza E, Aite M, Boutin M, Le Cam A, Montfort J, et al. Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. *BMC Genom.* (2018) 19:187. doi: 10.1186/s12864-018-4520-5

67. Chen Y, Zhao Y, Jin W, Li Y, Zhang Y, Ma X, et al. MicroRNAs and their regulatory networks in Chinese Gushi chicken abdominal adipose tissue during postnatal late development. *BMC Genom.* (2019) 20:778. doi: 10.1186/s12864-019-6094-2

68. Suzuki S, Kobayashi M, Murai A, Tsudzuki M, Ishikawa A. Characterization of growth, fat deposition, and lipid metabolism-related gene expression in lean and obese meat-type chickens. *J Poult Sci.* (2019) 56:101–11. doi: 10.2141/jpsa.0180064

69. Cogburn LA, Trakooljul N, Wang X, Ellestad LE, Porter TE. Transcriptome analyses of liver in newly-hatched chicks during the metabolic perturbation of fasting and re-feeding reveals THRSPA as the key lipogenic transcription factor. *BMC Genom.* (2020) 21:109. doi: 10.1186/s12864-020-6525-0

70. Pirany N, Bakrani Balani A, Hassanpour H, Mehraban H. Differential expression of genes implicated in liver lipid metabolism in broiler chickens differing in weight. *Br Poult Sci.* (2020) 61:10–6. doi: 10.1080/00071668.2019.1680802

71. Dhamad A, Zampiga M, Greene ES, Sirri F, Dridi S. Neuropeptide Y and its receptors are expressed in chicken skeletal muscle and regulate mitochondrial function. *Gen Comp Endocrinol.* (2021) 310:113798. doi: 10.1016/j.ygcen.2021.113798

72. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303

73. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019) 47:607–13. doi: 10.1093/nar/gky1131

74. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* (2021) 49:605–12. doi: 10.1093/nar/gkaa1074

75. Sam SA, Teel J, Tegge AN, Bharadwaj A, Murali TM. XTalkDB: a database of signaling pathway crosstalk. *Nucleic Acids Res.* (2017) 45:432–9. doi: 10.1093/nar/gkw1037

76. Gamage DG, Gunaratne A, Periyannan GR, Russell TG. Applicability of instability index for *in vitro* protein stability prediction. *Protein Pept Lett.* (2019) 26:339–47. doi: 10.2174/0929866526666190228144219

77. Mbaye MN, Hou Q, Basu S, Teheux F, Pucci F, Rooman M. A comprehensive computational study of amino acid interactions in membrane proteins. *Sci Rep.* (2019) 9:12043. doi: 10.1038/s41598-019-48541-2

78. Gun L, Yumiao R, Haixian P, Liang Z. Comprehensive analysis and comparison on the codon usage pattern of whole *Mycobacterium tuberculosis* coding genome from different area. *Biomed Res Int.* (2018) 2018:3574976. doi: 10.1155/2018/3574976

79. Suzuki H, Morton BR. Codon adaptation of plastid genes. *PLoS ONE.* (2016) 11:e0154306. doi: 10.1371/journal.pone.0154306

80. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, et al. A new and updated resource for codon usage tables. *BMC Bioinform.* (2017) 18:391. doi: 10.1186/s12859-017-1793-7

81. Bahiri-Elitzur S, Tuller T. Codon-based indices for modeling gene expression and transcript evolution. *Comput Struct Biotechnol J.* (2021) 19:2646–63. doi: 10.1016/j.csbj.2021.04.042

82. Gómez-Tamayo JC, Cordomí A, Olivella M, Mayol E, Fourmy D, Pardo L. Analysis of the interactions of sulfur-containing amino acids in membrane proteins. *Protein Sci.* (2016) 25:1517–24. doi: 10.1002/pro.2955

83. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* (2007) 8:R232. doi: 10.1186/gb-2007-8-11-r232

84. Anashkina AA, Petrushanko IY, Ziganshin RH, Orlov YL, Nekrasov AN. Entropy analysis of protein sequences reveals a hierarchical organization. *Entropy.* (2021) 23:1647. doi: 10.3390/e23121647

85. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts.* (2010) 1:271–83. doi: 10.1515/bmc.2010.022

86. Haddad Y, Adam V, Heger Z. Rotamer dynamics: analysis of rotamers in molecular dynamics simulations of proteins. *Biophys J.* (2019) 116:2062–72. doi: 10.1016/j.bpj.2019.04.017

87. Lovell SC, Davis IW, Arendall WB III, De Bakker PI, Word JM, et al. Structure validation by Calpha geometry: phi, psi and Cbeta deviation. *Proteins*. (2003) 50:437–50. doi: 10.1002/prot.10286

88. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci USA*. (2016) 113:6117–25. doi: 10.1073/pnas.1606724113

89. Rey B, Spée M, Belouze M, Girard A, Prost J, Roussel D, et al. Oxygen recovery up-regulates avian UCP and ANT in newly hatched ducklings. *J Comp Physiol B Biochem Syst Environ Physiol.* (2009) 180:239–46. doi: 10.1007/s00360-009-0409-6

90. Rey B, Roussel D, Romestaing C, Belouze M, Rouanet JL, Desplanches D, et al. Up-regulation of avian uncoupling protein in cold-acclimated and hyperthyroid ducklings prevents reactive oxygen species production by skeletal muscle mitochondria. *BMC Physiol.* (2010) 10:5. doi: 10.1186/1472-6793-10-5

91. Jenni-Eiermann S, Jenni L, Smith S, Costantini D. oxidative stress in endurance flight: an unconsidered factor in bird migration. *PLoS ONE.* (2014) 9:e97650. doi: 10.1371/journal.pone.0097650

92. Akbarian A, Michiels J, Degroote J, Majdeddin M, Golian A, De Smet S. Association between heat stress and oxidative stress in poultry; mitochondrial dysfunction and dietary interventions with phytochemicals. *J Anim Sci Biotechnol.* (2016) 7:37. doi: 10.1186/s40104-016-0097-5

93. Mailloux RJ, Seifert EL, Bouillaud F, Aguer C, Collins S, Harper M-E. Glutathionylation acts as a control switch for uncoupling proteins UCP2 and UCP3. *JBC.* (2011) 286:21865–75. doi: 10.1074/jbc.M111.240242

94. Ueda M, Watanabe K, Sato K, Akiba Y, Toyomizu M. Possible role for avPGC-1alpha in the control of expression of fiber type, along with avUCP and avANT mRNAs in the skeletal muscles of cold-exposed chickens. *FEBS Lett.* (2005) 579:11–7. doi: 10.1016/j.febslet.2004.11.039

95. Brannan KE, Helfrich KK, Flentke GR, Smith SM, Livingston KA, Jansen Van Rensburg C. Influence of incubation, diet, and sex on avian uncoupling protein expression and oxidative stress in market age broilers following exposure to acute heat stress. *Poult Sci.* (2022) 101:101748. doi: 10.1016/j.psj.2022.101748

96. Xu R, Yu C, Mao L, Jiang M, Gao L, Li M, et al. Antioxidant defense mechanisms and fatty acid catabolism in Red-billed Leiothrix (*Leiothrix lutea*) exposed to high temperatures. *Avian Re.* (2022) 13:100013. doi: 10.1016/j.avrs.2022.100013

97. Zhang L, Yue HY, Wu SG, Xu L, Zhang HJ, Yan HJ, et al. Transport stress in broilers. II. Superoxide production, adenosine phosphate concentrations, and mRNA levels of avian uncoupling protein, avian adenine nucleotide translocator, and avian peroxisome proliferator-activated receptor-gamma coactivator-1alpha in skeletal muscles. *Poult Sci.* (2010) 89:393–400. doi: 10.3382/ps.2009-00281

98. Li Z, Miao Z, Ding L, Teng X, Bao J. Energy metabolism disorder mediated ammonia gas-induced autophagy via AMPK/mTOR/ULK1-Beclin1 pathway in chicken livers. *Ecotoxicol Environ Saf.* (2021) 217:112219. doi: 10.1016/j.ecoenv.2021.112219

99. Vert G, Chory J. Crosstalk in cellular signaling: background noise or the real thing? *Dev Cell.* (2011) 21:985–91. doi: 10.1016/j.devcel.2011.11.006

100. Müller IE, Rubens JR, Jun T, Graham D, Xavier R, Lu TK. Gene networks that compensate for crosstalk with crosstalk. *Nat Commun.* (2019) 10:4028. doi: 10.1038/s41467-019-12021-y

101. Grah R, Friedlander T. The relation between crosstalk and gene regulation form revisited. *PLoS Comput Biol.* (2020) 16:e1007642. doi: 10.1371/journal.pcbi.1007642

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fvets.2023.1109188

Check for updates

OPEN ACCESS

EDITED BY
Carlos Eduardo Fonseca-Alves,
Paulista University, Brazil

REVIEWED BY
Stanislaw Dzimira,
Wrocław University of Environmental and Life Sciences, Poland
Cord Brundage,
University of Wisconsin–La Crosse,
United States

*CORRESPONDENCE
Gillian Dank
✉ gilliandank@gmail.com

†PRESENT ADDRESS
Gillian Dank,
Onco Pet, Kfar Saba, Israel

SPECIALTY SECTION
This article was submitted to Oncology in Veterinary Medicine, a section of the journal Frontiers in Veterinary Science

RECEIVED 27 November 2022
ACCEPTED 09 January 2023
PUBLISHED 26 January 2023

CITATION
Dank G, Buber T, Polliack G, Aviram G, Rice A, Yehudayoff A and Kent MS (2023) A pilot study for a non-invasive system for detection of malignancy in canine subcutaneous and cutaneous masses using machine learning. *Front. Vet. Sci.* 10:1109188.
doi: 10.3389/fvets.2023.1109188

A pilot study for a non-invasive system for detection of malignancy in canine subcutaneous and cutaneous masses using machine learning

Gillian Dank[1]*†, Tali Buber[2], Gabriel Polliack[2], Gal Aviram[2,3], Anna Rice[2], Amir Yehudayoff[4] and Michael S. Kent[5]

[1]Koret School of Veterinary Medicine, Hebrew University, Rehovot, Israel, [2]HT BioImaging LTD, Hod Hasharon, Israel, [3]Department Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel, [4]Department of Mathematics, Technion, Haifa, Israel, [5]Department of Surgical and Radiological Sciences, School of Veterinary Medicine, University of California, Davis, Davis, CA, United States

Introduction: Early diagnosis of cancer enhances treatment planning and improves prognosis. Many masses presenting to veterinary clinics are difficult to diagnose without using invasive, time-consuming, and costly tests. Our objective was to perform a preliminary proof-of-concept for the HT Vista device, a novel artificial intelligence-based thermal imaging system, developed and designed to differentiate benign from malignant, cutaneous and subcutaneous masses in dogs.

Methods: Forty-five dogs with a total of 69 masses were recruited. Each mass was clipped and heated by the HT Vista device. The heat emitted by the mass and its adjacent healthy tissue was automatically recorded using a built-in thermal camera. The thermal data from both areas were subsequently analyzed using an Artificial Intelligence algorithm. Cytology and/or biopsy results were later compared to the results obtained from the HT Vista system and used to train the algorithm. Validation was done using a "Leave One Out" cross-validation to determine the algorithm's performance.

Results: The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of the system were 90%, 93%, 88%, 83%, and 95%, respectively for all masses.

Conclusion: We propose that this novel system, with further development, could be used to provide a decision-support tool enabling clinicians to differentiate between benign lesions and those requiring additional diagnostics. Our study also provides a proof-of-concept for ongoing prospective trials for cancer diagnosis using advanced thermodynamics and machine learning procedures in companion dogs.

KEYWORDS

dogs, oncology, machine learning, diagnosis, artificial intelligence, neoplasia, screening test

Introduction

Cancer is the leading cause of death in 45–47% of dogs over 10 years of age (1, 2). Cancer diagnosis is of key importance in treatment planning and providing better treatment. The ability to easily diagnose early-stage neoplasia in general practices should improve prognosis dramatically. Currently, either fine-needle aspiration or biopsies are the recommended diagnostic tests for subcutaneous and cutaneous masses. In many cases, these procedures are easily performed; however, they may be highly invasive and costly, which can delay an owner's decision to pursue a diagnostic workup. Studies that have compared the diagnostic accuracy for the diagnosis of neoplasia of fine-needle aspirates to histopathologic results showed a negative predictive value (NPV) of 63.63 and 68.7% in reviewed series in dogs and cats (3, 4). These results demonstrate the need for an alternative non-invasive procedure for early cancer detection.

The HT Vista device is based on the differences between malignant and normal tissue properties, primarily the fact that both tissues display different heat transfer rates (5, 6). These thermophysical properties are affected by the differences between the compositions, morphology, and vascular networks of the tissues (7–10). The calculated rate at which heat transfers throughout a material is termed thermal diffusivity. This diffusivity is determined by three major properties: thermal conductivity, heat capacity, and density (11). In the case of living tissues, their metabolism and blood flow dramatically affect their heat transfer (12, 13). In well-established tumor tissue, characterized by increased metabolic activity, faster growth processes, and increased blood vessel generation and usage, an increase of roughly one degree Celsius, compared to healthy neighboring tissue, was reported (14). This further supports the premise that cancer cells have different thermal properties compared to normal tissues.

In this study, we hypothesize that thermal diffusivity will differ between malignant and benign canine subcutaneous and cutaneous masses and that the HT Vista algorithm would be able to differentiate these masses into either the benign or malignant categories.

Materials and methods

Study design and case collection

This is a prospective study that was approved by the ethical review board committee (HU-NER-2020-015-A). The study population included dogs that presented to the Veterinary Teaching Hospital at the Koret School of Veterinary Medicine (Rishon LeZion, Israel). Informed consent was obtained from the owners of all the dogs prior to enrollment in the study.

The inclusion criteria included a signed owner consent, an externally accessible subcutaneous or cutaneous mass or lymph node that could be palpated, measured, and imaged by the device and considered safe for the dog to undergo an aspirate or biopsy. Dogs were excluded from the study if they had no gross disease, the mass was inflamed or infected, or if the mass was larger than 15 cm. In addition, the case was excluded if the cytology or biopsy did not provide a diagnosis, or if the thermal imaging was not successful. All dogs were monitored for adverse effects.

The data acquired by the HT Vista system did not influence any subsequent treatment or decision-making, and the clinician and the pathology lab were blinded to the results obtained by the system. Demographic information, as well as tumor measurements and location, were recorded using standard manual case reporting forms.

The device

The HT Vista system (hereafter termed "the system") is based on a continuous measurement of heat diffusion through the tissue. The system is composed of a control unit which includes a mini personal computer with internet capabilities, a touch screen, a dedicated software application, and a handheld probe. The probe consists of an optical camera, a high-power LED (Light-Emitting Diode) emitter (i.e., the heating source), and an inherent LWIR (long-wave-infra-red) thermal video camera, which records the temperature throughout the scan.

Patient preparation and testing process

The dogs were manually restrained, and the mass area was clipped. The probe was positioned above the examined area, which was subsequently scanned. The scan lasted 60 s. This included both heating of the target area by the high-power LED emitter by seven degrees Celsius for 10 s and continuous recording of the heat emitted by the tissue during and post-heating by the LWIR video camera. Then, the clinician marked two areas on an optical image of the scanned area, presented on the touch screen. The first selection represented the mass area (i.e., "site"), while the second one, adjacent to the mass, represented a normal tissue (i.e., "control"). If there were areas with different pigmentation, areas with the same pigmentation were marked. Then, unique thermal signals were produced, showing the changes in temperature in the site and the control throughout the test, based on the selection of healthy and suspicious sites by the clinician. The data obtained were uploaded to the HT Bioimaging cloud and analyzed using signal analysis techniques and a dedicated HT machine learning algorithm. The clinician was blinded to the results. Finally, the tested mass was aspirated and/or biopsied, according to the clinical recommendations. The aspirates were performed with a 25 gauge needle and submitted to an external pathology laboratory. The biopsies were performed by the surgery department and submitted to an external laboratory. Both the clinical pathologists and anatomic pathologists examining the samples were blinded to the results.

Dataset description

Both marked sites were presented as areas of 1.5 X 1.5 mm^2. Each was composed of a 25 pixels grid. The thermal signal of each of the pixels in both sites was represented by a set of ca. 1,000 signal descriptive features (i.e., values). These extracted features were based on mathematical, physical, and thermal properties, such as coefficients of the Pennes equation, as well as properties derived from signal analysis [e.g., Fourier series coefficients, used to describe periodic signals (15)]. The features of the control site pixels were integrated into the features of the mass site pixels, ensuring that the differences between the tissues were considered. All features were normalized to eliminate possible variances between patients and anatomical areas. Next, the control site pixels were removed from the dataset, resulting in a dataset of >1,000 normalized feature values, for each of the 25 mass site pixels, per patient. Finally, the results obtained from the cytology and/or histopathology of the mass were used to label the site as malignant or benign for subsequent training of the HT Vista algorithm.

Training and validation procedure

The final training procedure was performed on a set of the four most important features that best differentiated between benign and malignant lesions, including two Fourier Series coefficients and a fitted decay function coefficient. Sites were labeled as malignant or benign, as described above. The data were trained using a Support Vector Machine (SVM) classifier, a widely used AI classification algorithm. The training and validation were done using a "Leave-One-Out" cross-validation procedure to demonstrate that the classifier represents a general pattern. Cross-validation is a well-established practice in machine learning for model performance evaluation on limited data. This procedure partitions the data into *N* subsets, iteratively training the classifier on *N-1* different subsets in each iteration. Then, it uses the one left-out subset as a test set (i.e., classifies the single subset it was not trained on as malignant or benign). In this procedure, all instances are eventually used as both training and test sets. Commonly, multiple iterations of cross-validation are performed, and performance assessments are averaged over all iterations to increase robustness and reduce variability (16–18). Specifically, the "Leave-One-Out" cross-validation withholds in each iteration a single set of pixels belonging to the same mass (i.e., the same patient), while the other masses and their pathology results train the classifier. The trained classifier was then applied to the data of the one mass left out, resulting in a classification of the marked tested area as either high-risk or low-risk (i.e., malignant or benign).

After classifying all lesions, the performance of the algorithm was assessed using a confusion matrix. The matrix summarizes the identities and differences between the real diagnosis obtained from cytology or histopathology and the predictions made by the algorithm. Each cell of the matrix holds the number of correct and incorrect classifications made by the algorithm of each of the possible classes. That is, the matrix counts how many true-positives (malignant lesions classified as malignant), true-negative (benign lesions classified as benign), false-positive (benign lesions classified as malignant), and false-negative (malignant lesions classified as benign) cases were found in the study. The overall performance of the HT algorithm was then assessed by calculating five measures: (7) Accuracy–the overall fraction of correct classifications. (1) Sensitivity- the fraction of high-risk predicted lesions within the malignant or premalignant pathology reports. (3) Specificity–the fraction of low-risk predicted lesions within the benign or non-malignant pathology reports. (4) Positive predictive value (PPV)–the fraction of true positives within the high-risk predictions. (5) Negative predictive value (NPV)–the fraction of true negatives within the low-risk predictions.

Results

Forty-nine dogs were initially included in the study. A total of four dogs were excluded: one mass was not diagnosed, one was not sufficiently clipped for the scan to be diagnostic, and two did not have healthy areas imaged during the scan. A final group of 45 dogs met the inclusion criteria. Thirty-three were mixed-breed dogs, and 12 were

TABLE 1 Classification of cytology and histopathology results.

Tumor type				
Benign		Cytology	Histopathology	Total
	Adenoma	1	3	4
	Benign epithelial/adnexal cyst/tumor	4	1	5
	Benign melanoma		1	1
	Lipoma	20	1	21
	Papilloma	1		1
	Perineal adenoma	1	1	2
	Plasmacytoma		1	1
	Reactive lymph node	2		2
	Sebaceous adenoma	3		3
	Sebaceous hamartoma		1	1
	Sebaceous hyperplasia	1		1
Benign Total		33	9	42
Malignant		Cytology	Histopathology	Total
	Adenocarcinoma		3	3
	Cutaneous hemangiosarcoma		3	3
	Lymphoma	10		10
	Mast cell tumor	3	2	5
	Metastatic lymph node (hemangiosarcoma)	1		1
	Metastatic lymph node (melanoma)		1	1
	Osteosarcoma		1	1
	Soft tissue sarcoma		1	1
	Undifferentiated neoplasia		2	2
Malignant total		14	13	27

purebred dogs. No purebred dog was over-represented. There were 16 intact female dogs, three spayed female dogs, and 26 intact male dogs. The median age was 11 years, ranging between four and 14.

Of the 45 dogs, 24 had one lesion sampled, 18 dogs had two lesions sampled, and three dogs had three lesions sampled, resulting in a total of 69 lesions. Twenty dogs were classified with 27 malignant lesions based on their cytology or histopathological diagnosis. Forty-eight lesions were diagnosed using cytology, and 21 lesions were diagnosed using histopathology (Table 1).

Using the machine learning classifier, each examined site was classified as either high-risk or low-risk for malignancy, and results were compared to the pathology reports. These results are shown in Table 2. In total, 62 out of 69 lesions were correctly classified, 25 as malignant and 37 as benign, while seven were misclassified. Five were false-positive classifications, including one keratinous cyst, three deep lipomas, and one sebaceous gland adenoma. The other two were false-negative cases which included two lymphomas. The overall accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were 90, 93, 88, 83, and 95%, respectively.

Discussion

AI-driven medical devices are becoming more and more common in veterinary medicine. They are used to solve problems of high logical or algorithmic complexity, ranging from diagnosis and disease detection to making reliable predictions and reducing medical errors (19). In this study, we introduced a novel diagnostic AI-based imaging system, the HT Vista, which aims to provide a high degree of accuracy in differentiating between benign and malignant, cutaneous and subcutaneous lesions in dogs, based on the response of a tissue to thermal excitation. Malignant tumor tissues differ from normal tissues by their known high metabolic rate and increased perfusion and their capability to transfer heat, which in turn shows a different response to thermal excitation (11–14). Additional factors that have been reported to influence thermal imaging in dogs include inflammation, infection, trauma and temperature at the time of the scan (20). However, our results do not appear to have been influenced by these factors. We found that the temperatures recorded during the thermal relaxation phase distinguished between normal and malignant tissues. An AI-based algorithm was trained on physical

TABLE 2 Results of the HT Vista system classification for all sites.

		HT classification		
Benign/malignant	Tumor type	Benign	Malignant	Total
	Adenoma	4		4
	Benign epithelial/adnexal cyst/tumor	4	1	5
	Benign melanoma	1		1
	Lipoma	18	3	21
	Papilloma	1		1
	Perineal adenoma	2		2
	Plasmacytoma	1		1
	Reactive lymph node	2		2
	Sebaceous adenoma	2	1	3
	Sebaceous hamartoma	1		1
	Sebaceous hyperplasia	1		1
Benign total		37	5	42
		Benign	**Malignant**	**Total**
	Adenocarcinoma		3	3
	Cutaneous hemangiosarcoma		3	3
	Lymphoma	2	8	10
	Mast cell tumor		5	5
	Metastatic lymph node (hemangiosarcoma)		1	1
	Metastatic lymph node (melanoma)		1	1
	Osteosarcoma		1	1
	Soft tissue sarcoma		1	1
	Undifferentiated neoplasia		2	2
Malignant total		2	25	27

and thermal features, using samples labeled as malignant or benign based on the blinded pathological results.

Our study included a wide range of both benign and malignant tumors. The overall accuracy of the system was 90%, correctly classifying 62 out of 69 masses, with only two false negatives (lymph nodes diagnosed with lymphoma). The explanation for the misclassification of these cases was most likely the different anatomic structures of lymph nodes, which might have resulted in inadequate heating to elicit a representative thermal signal.

Five were false-positive classifications, including one keratinous cyst, three deep lipomas, and one sebaceous gland adenoma. Several explanations for these false results include that deep inhabiting tumors may require a change of the heat source configuration (e.g., the wavelength and penetration characteristics) and that the liquid content within cysts may heat differently. In any case, positive results should cause the clinician to continue to diagnose the mass, which will lead the clinician to conclude that this is a benign mass in cases of false positives. This is preferable to a higher number of false negatives, which would cause clinicians to send home animals with malignant tumors.

The HT Vista system's algorithm was programmed to give a high degree of certainty in classifying a mass as benign, thus minimizing the risk of false-negative cases. Therefore, in high-risk cases, this system enables the clinician to recommend continuing to work up these masses with either an aspiration and /or a biopsy and not take the wait-and-see approach.

Factors that influence cancer detection include inflammation and infection, which can cause dysplasia and lead to false positive results on the fine needle aspirates. In this study, there was one case of sebaceous hyperplasia that was a true negative based on the classifier. Additional causes that may influence detection include acellular samples, as can occur in cases of lipomas and sarcomas. In this study, the device accurately classified 18/21 lipomas and all carcinomas and sarcomas.

In this study, the performance of the algorithm's classifier was assessed using a "Leave-One-Out" cross-validation method, which is cross-validation taken to its extreme. This method is useful for evaluating machine learning models with a limited data set, as in our study, and provides an accurate and unbiased estimate of model performance (21).

The limitations of the study include the low number of cases and that the deeper tumors, including both deep lipomas and lymph nodes, may require changing the heat source configuration, as previously mentioned. In addition, cutaneous and epidermal tumors do not always present the same way or have the same disease progression as mesenchymal tumor types, which may cause a variation in the thermal signal. Skin pigmentation was not shown to have an effect on the thermal heating in this study, however, should be further assessed in future studies with additional dogs with different pigmentation. Additionally, larger studies should help give an understanding whether these differences affect thermal diffusivity.

Future directions include an additional multi-center trial with a larger study population to validate the system. As machine learning accuracy improves with additional data, the HT Vista's algorithm is expected to improve its analytic capabilities. Therefore, it is expected to provide more accurate results in the future.

In conclusion, in this study, we showed a proof of concept of a novel non-invasive diagnostic method and decision support tool for the clinical management of cutaneous and subcutaneous masses in dogs, using dynamic heat diffusivity and analysis of the produced signal utilizing advanced machine learning.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The animal study was reviewed and approved by Hebrew University HU-NER-2020-015-A. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

All authors contributed to the study-including the study planning, case accrual, algorithm, and writing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

HT Bioimaging fully funded this clinical trial that was performed at the Koret School of Veterinary Medicine, Hebrew University.

Acknowledgments

We wish to thank Moshe Tshuva, Sharon Gat, Assaf Gur, and Hen Toledano for their help with the development of the device.

Conflict of interest

GD was on the HT Bioimaging Scientific Advisory Board at the time of the study and is now employed by HTVet. MK was on the HT Bioimaging Scientific Advisory Board. TB, GP, GA, AR, and AY are employed by HTVet.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Bronson RT. Variation in age at death of dogs of different sexes and breeds. *Am J Vet Res.* (1982) 43:2057–9.

2. *Withrow and MacEwen's Small Animal Clinical Oncology—5th Edition*. (2012) Available online at: https://www.elsevier.com/books/withrow-and-macewens-small-animal-clinical-oncology/9781437723625 (accessed January 19, 2022).

3. Ghisleni G, Roccabianca P, Ceruti R, Stefanello D, Bertazzolo W, Bonfanti U, et al. Correlation between fine-needle aspiration cytology and histopathology in the evaluation of cutaneous and subcutaneous masses from dogs and cats. *Vet Clin Pathol.* (2006) 35:24–30. doi: 10.1111/j.1939-165X.2006.tb00084.x

4. Simeonov RS. The accuracy of fine-needle aspiration cytology in the diagnosis of canine skin and subcutaneous masses. *Comp Clin Path.* (2012) 21:143–7. doi: 10.1007/s00580-010-1075-5

5. Stefanadis C, Chrysohoou C, Panagiotakos DB, Passalidou E, Katsi V, Polychronopoulos V, et al. Temperature differences are associated with malignancy on lung lesions: a clinical study. *BMC Cancer.* (2003) 3:1–5. doi: 10.1186/1471-2407-3-1

6. Yates AJ, Thompson DK, Boesel CP, Albrightson C, Hart RW. Lipid composition of human neural tumors. *J Lipid Res.* (1979) 20:428–36. doi: 10.1016/S0022-2275(20) 40596-6

7. Ho B, Kannayiram K, Tam R, Yang H. *Modeling temperature in a breast cancer tumor for ultrasound-based hyperthermia treatment* (Thesis). University of California. (2012).

8. Holmes KR, Ryan W, Chen MM. Thermal conductivity and H2O content in rabbit kidney cortex and medulla. *J Therm Biol.* (1983) 8:311–3. doi: 10.1016/0306-4565(83)90014-1

9. Welch AJ, Van Gemert MJC. *Optical-Thermal Response of Laser-Irradiated Tissue*. New York, NY: Springer (1995).

10. Storm FK, Harrison WH, Elliott RS, Morton DL. Normal tissue and solid tumor effects of hyperthermia in animal models and clinical trials. *Cancer Res.* (1979) 39: 2245–51.

11. Jiji Latif M. Heat conduction: Third edition. *Heat Conduction.* (2009) 1–418. doi: 10.1007/978-3-642-01267-9_1

12. Minkowycz WJ. *Advances in Numerical Heat Transfer, Volume 3*. Boca Raton, FL: CRC Press (2009).

13. Pennes HH. Analysis of tissue and arterial blood temperatures in the resting human forearm. *J Appl Physiol.* (1948) 1:93–122. doi: 10.1152/jappl.1948.1.2.93

14. Papetti M, Herman IM. Mechanisms of normal and tumor-derived angiogenesis. *Am J Physiol Cell Physiol.* (2002) 282:C947–70. doi: 10.1152/ajpcell.00389.2001

15. Bracewell RN. *The Fourier Transform and its Applications*. Toledo, OH: Discover (1986). p. 474.

16. Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Review of medical decision support and machine-learning methods. *Vet Pathol.* (2019) 56:512–25. doi: 10.1177/0300985819829524

17. Dietterich TG. Machine-learning research. *AI Mag.* (1997) 18:97–97.

18. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *The International Joint Conference on Artificial Intelligence*. (1995).

19. Ezanno P, Picault S, Beaunée G, Bailly X, Muñoz F, Duboz R, et al. Research perspectives on animal health in the era of artificial intelligence. *Vet Res.* (2021) 52:40. doi: 10.1186/s13567-021-00902-4

20. Sung J, Loughin C, Marino D, Leyva F, Dewey C, Umbaugh S, et al. Medical infrared thermal imaging of canine appendicular bone neoplasia. *BMC Vet Res.* (2019) 15:430. doi: 10.1186/s12917-019-2180-6

21. Raschka S, Mirjalili V. *Python Machine Learning: Machine Learning and Deep Learning With Python, Scikit-Learn, and Tensorflow*. Birmingham: Packt Publishing Ltd (2019).

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 21 February 2023
DOI 10.3389/fvets.2023.1128570

Benchmark study for evaluating the quality of reference genomes and gene annotations in 114 species

Sinwoo Park[1†], Jinbaek Lee[2†], Jaeryeong Kim[1], Dohyeon Kim[1], Jin Hyup Lee[3], Seung Pil Pack[4] and Minseok Seo[1,2]*

[1]Department of Computer and Information Science, Korea University, Sejong City, Republic of Korea, [2]Department of Computer Convergence Software, Korea University, Sejong City, Republic of Korea, [3]Department of Food and Biotechnology, Korea University, Sejong City, Republic of Korea, [4]Department of Biotechnology and Bioinformatics, Korea University, Sejong City, Republic of Korea

OPEN ACCESS

EDITED BY
Rajesh Kumar Pathak,
Chung-Ang University, Republic of Korea

REVIEWED BY
Jang-il Sohn,
Genome4me, Republic of Korea
Donghyun Shin,
Jeonbuk National University, Republic of Korea

*CORRESPONDENCE
Minseok Seo
✉ mins@korea.ac.kr

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

RECEIVED 20 December 2022
ACCEPTED 02 February 2023
PUBLISHED 21 February 2023

CITATION
Park S, Lee J, Kim J, Kim D, Lee JH, Pack SP and Seo M (2023) Benchmark study for evaluating the quality of reference genomes and gene annotations in 114 species. *Front. Vet. Sci.* 10:1128570. doi: 10.3389/fvets.2023.1128570

Introduction: For reference genomes and gene annotations are key materials that can determine the limits of the molecular biology research of a species; however, systematic research on their quality assessment remains insufficient.

Methods: We collected reference assemblies, gene annotations, and 3,420 RNA-sequencing (RNA-seq) data from 114 species and selected effective indicators to simultaneously evaluate the reference genome quality of various species, including statistics that can be obtained empirically during the mapping process of short reads. Furthermore, we newly presented and applied transcript diversity and quantification success rates that can relatively evaluate the quality of gene annotations of various species. Finally, we proposed a next-generation sequencing (NGS) applicability index by integrating a total of 10 effective indicators that can evaluate the genome and gene annotation of a specific species.

Results and discussion: Based on these effective evaluation indicators, we successfully evaluated and demonstrated the relative accessibility of NGS applications in all species, which will directly contribute to determining the technological boundaries in each species. Simultaneously, we expect that it will be a key indicator to examine the direction of future development through relative quality evaluation of genomes and gene annotations in each species, including countless organisms whose genomes and gene annotations will be constructed in the future.

KEYWORDS

reference genome, gene annotation, quality assessment, transcript diversity, next-generation sequencing (NGS), RNA-sequencing (RNA-seq), livestock animals, model organisms

Introduction

Next-generation sequencing (NGS) technology is applied in many ways to identify the biological characteristics of various organisms, including livestock, at the molecular level (1, 2). This technology is used in virtually all biomedical fields, such as research to find genetic variants based on DNA sequencing (3, 4) and research to discover transcripts related to life phenomena based on RNA-sequencing (RNA-seq) (5, 6). Recently, NGS technology has been developed for data acquisition of molecular characteristics at the level of single cells (7) or single nuclei (8), concurrently, long-read-based technologies are continuously being developed to improve sequencing quality (9). Various technologies are continuously being developed to measure various levels of molecular markers more accurately; however, all of them are strongly dependent on the reference genome and gene annotation corresponding to the biological species of the targeted subject in certain studies (10). As of 2023, the fundamental and essential data of the NGS technique, reference genomes and gene annotations, have been established in the Ensembl database for 314 species (11). Moreover, it is highly likely that the number of completed reference genomes and gene annotations for more species will increase exponentially in the near future through the vertebrate genome project (VGP) (12). Thus, a relative comparison of relevant essential data is necessary to increase the reliability of various applied studies in more diverse species. Although the accuracy of the results of each study utilizing NGS highly depends on the completeness of the two key underlying data, there has been no systematic evaluation of reference genomes and gene annotations among diverse species simultaneously. Although, species have a common genetic background, to some extent, the genome structure, number, and type of transcripts differ considerably between organisms, which makes comparisons across species quite challenging (13, 14).

To date, various attempts have been made to identify the whole-genome sequence in a particular species by selecting the optimal assembly from a number of draft assemblies. Various methodologies such as, KAT (15), Merqury (16), and Inspector (17), have been developed to compare the quality of different versions of draft assemblies for a specific target species to determine a representative genome. However, these methodologies require whole genome sequencing (WGS) reads and/or a reference genome of the target species, therefore, they cannot be directly applied for the purpose of evaluating the quality of reference genomes for multiple species. Among these tools, BUSCO (18, 19) can be used to compare the quality of reference genomes for multiple species based on the orthologous genes. However, since the optimal assembly was already determined in the direction of optimizing the BUSCO completeness in the process of completing the reference assembly of each species, the difference in BUSCO completeness of the published reference genome is very small among species. Although we currently lack systematic methodologies that can be used to directly and simultaneously compare the quality of reference assemblies of various species, some indicators can be used to compare species. First, the quality of the reference genome was compared using a contiguity index, such as the N50 value obtained based on the relative length of contigs or scaffolds generated during the *de novo* assembly process (20–22). Another quality evaluation index for the completed genome is the number and frequency of gaps in the genome, and various attempts have been made to reduce them (23–25). However, gene annotation quality assessment methods remain poorly understood, owing to their transcriptome diversity. Recently, software has been developed that can estimate the annotation similarity of evolutionarily adjacent species based on the gene annotations of species known to be nearly complete, allowing a relative comparison of the gene annotations of the two species (26). However, there is, to date, no known systematic approach to compare gene annotations of multiple species.

Although long-read sequencing technology is continuously being refined, NGS application research is still mainly based on short-read sequencing technology. RNA-seq, a representative application of NGS based on short reads, generally involves a two-step analysis. The first step is an alignment process to determine where the short-read fragmented sequences originate from the genome, for which the quality of the reference genome is important (27, 28). If the accuracy of the sequence of the reference genome is low, the mapping rate is directly affected. If the frequency of repeat sequences is high, the number of multiple mapping reads increases, adversely affecting the entire process. The second major step for processing RNA-seq data is to quantify the mapped reads in the genome (29). At this time, performance greatly depends on the quality of the gene annotation, which defines the location of the transcripts in the genome (30, 31). If all transcripts that can occur in a specific organism are included in gene annotation, the quantification rate will increase; however, the probability of overlapping other transcripts at a specific genome location will correspondingly increase, resulting in quantification failure due to ambiguity. Concurrently, inclusion of transcripts that are too conservative in gene annotations to address this ambiguity exacerbates quantification failures caused by the absence of annotations. These issues are commonly considered when developing reference genomes and gene annotations for various species, thus the quality of the two fundamental types of data can be measured indirectly through the corresponding indicators at the alignment and quantification steps.

Based on these rationales, in this study, we attempted to evaluate the quality of reference genomes and gene annotations of all species as much as possible, which has not yet been performed because of technical issues. We attempted to measure the quality of two key data essential in NGS from various angles by assessing the effectiveness of new potential indicators along with the indicators that have been used so far for quality evaluation. In addition, we aimed to demonstrate a new integrated index for the simultaneous quality evaluation of genome and gene annotation, by applying selected quality effective indicators to RNA-seq data derived from various species.

Materials and methods

Reference genome and gene annotation collection

As of November 2022, the latest genome assembly (.fasta) of each species and the corresponding gene annotation (.gtf) were collected from the Ensembl database (Supplementary Table 1) using Rcurl v1.98.1. Among all species, human, mouse, and zebrafish species that had access to the primary assembly version were used, and the toplevel version of the genome was used for the rest of the species.

Collection of basic statistic on genome assembly and gene annotation

Basic assembly information for all species was collected in xml format through the API of ENA (European Nucleotide Archive) (https://www.ebi.ac.uk/ena/browser/api/xml/Assembly accession). The collected assembly basic statistics were tabulated using xml2 (v1.3.3) and tidyverse (v1.3.2) R packages. We also collected detailed information on gene annotation from Ensembl biomart (32) using the biomaRt (v2.50.3) R package. Using the getBM function, various information including ensemble gene id and gene type were collected and tabulated from the gene annotation of each species. The transcript types in gene annotation were classified into 30 types according to the classification criteria of Ensembl gene biotype (https://asia.ensembl.org/info/genome/genebuild/biotypes.html) (Supplementary Table 2).

Estimation of repeat elements from reference genomes

The Repeat Masker (v4.1.4) (33) with -pa 16 -qq options was used to quantify repeat elements from reference genomes of various species. RMBlast (v2.11.0) was used as the repetitive sequence search algorithm, and the search was based on the Dfam (v3.6) database (34). In addition, TRF (v4.09) (35) was used to find tandem repeat sequences.

RNA-seq raw data collection

As of November 2022, among the species whose reference genome and gene annotation are listed in the Ensembl database, we searched for species that could secure RNA-seq data of more than 30 samples. Using R (v4.1.2) language-based packages XML (v3.99.0.12) and xml2 (v1.3.3), data corresponding to the following conditions was retrieved from NCBI Esearch (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi) and 30 SRA IDs of each species were randomly selected. In XML parsing with the GET method, we consider the following four conditions: “biomol rna”, “library layout paired”, “platform illumina”, and “Bulk”. After that, we used the prefetch (v2.11.2) included in the SRAtoolkit (v2.11.3) to import randomly selected sra files from the SRA database (36). To convert the collected sra files into paired-end fastq format files, parallel-fastq-dump was employed. The FastQC v.0.11.9 (37) was used to check the quality of the collected raw sequencing data.

Preprocessing of RNA-seq data

All collected genomes were indexed using the full Hisat2-build (v.2.2.1) (38). Paired-end RNA-seq files whose quality was checked through FastQC (v.0.11.9) were mapped to each corresponding genome. Alignment results were recorded in sorted bam format through samtools view (v1.14), and mapping-related statistics were collected through samtools stats. The mapped reads to each genome were quantified using featureCounts (v2.0.1) (39) with the corresponding gene annotation.

Quality evaluation indicators for reference genome in diverse species

A total of 10 indicators used in this study are summarized in Table 1. All indicators are scaled in the range of 0–1 for the convenience of interpretation. Also, the closer the value is to 1, showing the better the quality in all indicators.

As indicators for simultaneous relative evaluation of the genomes of various species, three indicators were selected based on the statistics derived from the assembly process. Based on the N50 values of contig and scaffold, which are the continuity indices of assembly, it was corrected to consider the different genome size of various species. These corrected N50s were converted to have a range of 0 to 1 by their percentile. Through this, two indicators, *AdjN50Contig* and *AdjN50Scaffold*, were calculated respectively. Next, to get the *UngapRate*, it was subtracted from 1 to adjust the directionality after obtaining the ratio of spanned gaps in the genome of each species compared to the species with the largest spanned gaps among all species.

We selected three empirical indicators obtained through the process of mapping actual NGS data as another measure to evaluate the quality of the genome. First, *UnimapRate* is basically the most important indicator in the mapping step, and represents the ratio of reads uniquely mapped to a specific genomic region among all reads. In addition, we additionally considered the two typical causes of mapping failure: multi-region mapping and no corresponding region. To match the direction as a quality evaluation index, *MapRate* and *MultiMapRate* indexes were constructed by subtracting the two failure rates from 1, respectively. Based on these three empirical indicators, we construct a new mapping quality evaluation index (MQI) for species *i*:

$$MQI_i = (UnimapRate_{\,i} + MapRate_{\,i} + MultiMapRate_{\,i}) / 3 \quad (1)$$

The MQI_i is the arithmetic mean of the three different directional indices obtained empirically from the mapping step, and is a relatively comparable indices across different species. Additionally, the BUCSO completeness was calculated using BUSCO (v5.4.2) with–auto—lineage-euk–cpu 16 options (18).

Quality evaluation indicators for gene annotation in diverse species

To qualitatively evaluate the quality of gene annotation, the proportion of each gene type was calculated based on the gene types collected from Ensembl biomart (32). Based on a matrix with a total of p gene type ratios for all species n, principal component analysis (PCA) was applied that can secure a linear combination of p gene type ratio random variables to convert to a nx1 vector for comparing all species n. After examining the degree of the variance explain based on the eigen values, the PC1 embedding values were extracted and used as *Transcript diversity*. Additionally, to further clarify the interpretation of PC1, another method of summarizing variability, Shannon’s equability index (40, 41), was calculated and compared.

As another criterion for evaluating the quality of the gene model, we selected three empirical indicators obtained through the process of quantifying reads mapped to the genome based on actual NGS data. First, *Quant.rate*, which is the ratio of reads successfully quantified as gene counts among mapped reads derived from each sample, was selected with the highest priority. Simultaneously, the absence and ambiguity of annotation, which are two representative quantification failure rate factors that can

TABLE 1 Selected 10 indicators for quality evaluation of reference genome and gene annotation in diverse species.

No.	Abbreviation of indicators	Description	Category	Formula	Scaling method	Range of values
1	AdjN50Contig	Percentile of adjusted N50 by genome size in contig	Assembly stat.	N50 value in contigs/genome size	Percentile	0–1
2	AdjN50Scaffold	Percentile of adjusted N50 by genome in scaffold	Assembly stat.	N50 value in scaffolds/genome size	Percentile	0–1
3	UngapRate	Scaled non-spanned gaps rate	Assembly stat.	1—[spanned gaps/max (spanned gaps)]	NA	0–1
4	UnimapRate	Uniquely mapped rate	Mapping stat.	Uniquely mapped reads/total # of reads	NA	0–1
5	MapRate	1—unmapped reads' rate	Mapping stat.	1—(unmapped reads/total # of reads)	NA	0–1
6	MultiMapRate	1—multiple mapped rate	Mapping stat.	1—(multiple mapped reads/total # of reads)	NA	0–1
7	Transcript diversity	Scaled transcript diversity calculated by PCA	Gene annotation	PC1 obtained from PCA analysis	{X—min (X)}/{max(X)—min (X)}	0–1
8	Quant.rate	Quantification success rate from the mapped reads on the genome	Quantification stat.	Quantification success reads/total # of mapped reads	NA	0–1
9	Quant.rate (Abs)	1—quantification failure rate due to absence of annotation	Quantification stat.	1—(unquantified mapped reads due to absence of annotation/total # of mapped reads)	NA	0–1
10	Quant.rate (Amb)	1—quantification failure rate due to ambiguity	Quantification stat.	1—(unquantified mapped reads due to ambiguity / total # of mapped reads)	NA	0–1

All indicators are scaled to have a value between 0 and 1 and the closer each index value is to 1 represents the better quality.

be determined by the gene model, were additionally considered. To match the directionality, two indicators, *Quant.rate (Abs)* and *Quant.rate (Amb)*, were set by subtracting the two failure rates from 1. Based on the three empirical indices obtained during the quantification process, we constructed the comprehensive quantification quality evaluation index (QQI) for species *i*:

$$QQI_i = (Quant.rate_i + Quant.rate(Abs)_i + Quant.rate(Amb)_i) / 3 \quad (2)$$

The QQI_i is the average of the three indices obtained empirically in the quantification stage of NGS data and is an indicator that can simultaneously compare the general quality of gene models in multiple species.

NGS applicability index

Based on a total of 10 effective indicators that can evaluate the genome and gene model (Table 1), it was generalized as an index representing the technical boundary of NGS technology in a specific species. The formula consisting of the weighted arithmetic mean of the 10 indicators for each species *i* is:

$$NGS\ applicability\ index_i = \frac{w_1 AdjN50Contig_i + w_2 AdjN50Scaffold_i + \ldots + w_{10} Quant.rate(Amb)_i}{\sum_{i=1}^{10} w_i} \quad (3)$$

In this study, all 10 weights w1, w2, ..., w10 were considered as 1, which means that all indicators are considered equally.

Results

Large-scale NGS data collection for quality evaluation of reference genomes and gene annotations of 114 species

We systematically collected data to evaluate the current level of reference genomes and gene annotations for as many species as possible, for which RNA-seq, among various NGS technologies, could be directly applied (Figure 1A). There were more than 30 publicly available RNA-seq datasets for 114 of the 314 species (Supplementary Table 1), whose reference genomes and gene annotations are listed in the Ensemble database (11). As a result of organizing the taxonomic categories for these 114 species compared in this study, it was confirmed that 1 fungus, 112 Metazoa, and 1 Viridiplantae were included at the kingdom level (Supplementary Table 3). At the taxonomic level, they were classified into 11 types, of which 47 Actinopteri, 43 Mammalia, and 12 Aves were the majority. In addition to collecting the latest version of the reference genome and gene annotation for these 114 species, 30 RNA-seq datasets per species were randomly collected, resulting in a total of 3,420 RNA-seq datasets (Supplementary Table 4). After the quality check, an average of 34 million reads and an average Phred score of 36.237 were observed, showing no technical issues in the collected RNA-seq data (Supplementary Table 5). When the collected RNA-seq data were mapped based on the reference genome representing each species, an average overall alignment rate of 84.768% was obtained (Supplementary Table 6). In quantification step, 55.807% of mapped reads were successfully quantified to genes in average (Supplementary Table 7).

To independently compare the quality of all 114 collected reference genomes, genome assembly statistics were compiled from the European Nucleotide Archive (42) and the corresponding information was missing for five species. The remaining 109 available species were systematically collected, and assembly related statistics were obtained from the collected data, resulting in an average length of 1,689,594,967 bp and a contig average N50 of 7,154,707 bp (Supplementary Table 8). We also collected data from Ensembl Biomart (32) to evaluate the quality of gene annotations that indicated the location of genic regions in the reference genome of each species; however, the information could not be collected for 12 out of 114. For the remaining 102 species, gene annotation was collected and classified as a total of 30 types of RNAs, including long non-coding RNA (lncRNAs) and microRNAs (miRNAs) (Supplementary Table 9). We found that an average of 22,915 protein-coding genes were annotated across all 102 species, while a significantly small number of average 2,340 lncRNAs were not annotated in 37 species.

Based on the collected data at various levels, an experimental design was established that measures the quality of the genomes and gene models in various species (Figure 1B). In this current study, we focused on quality measures for eight species of livestock designated according to the Food and Agriculture Organization of the United Nations (FAO).

Comparison of assembly statistics and frequency of repeat elements for reference genome quality evaluation in 109 species

Officially published reference genomes of various species are generally expected to show minimal difference in quality owing to the robustness of DNA; however, limitations exist due to the frequency of repetitive sequences in the genome and/or sequencing technology based on short reads. To investigate this, we collected and compared representative quality statistics of 109 genome assemblies, which were largely clustered into four characteristics (Figure 2A). While an average of 37,580.454 spanned gaps were found in all species, only 204 and 661 spanned gaps were found in the human and mouse genomes, respectively, which are known to be of high quality (Figure 2B). In addition, a significantly lower number of spanned gaps was observed in representative model organisms such as *Saccharomyces cerevisiae* (*S.cerevisiae*), *Arabidopsis thaliana* (*A.thaliana*), and *Drosophila melanogaster* (*D.melanogaster*) (Figure 2B, Supplementary Figure 1). While a low number of spanned gaps was found in most of the eight livestock animals, it was confirmed that a relatively large number of spanned gaps were present in the genomes of *Ovis aries* (125,067 gaps) and *Equus caballus* (6,286 gaps).

Furthermore, we found that the number of spanned gaps was strongly correlated with the number of contigs generated during the *de novo* assembly process, which revealed that in the case of species with many spanned gaps, relatively short contigs occurred during the assembly process (Figure 2A, Supplementary Figure 2). In other words, various technical issues derived from short sequence read assembly intensify depending on the number of spanned gaps ultimately affecting the quality of the completed genome assembly, which suggests that the genome quality of various species can be evaluated based on these statistics. Further evidence for this claim can be found in the negative correlations between the number of spanned gaps and adjusted N50, N75, and N90 values by genome size in both contigs and scaffolds (Figure 2A). These values are representative indicators used when evaluating the quality of the genome completed through *de novo* assembly, and significantly higher values were observed in representative model organisms at both the scaffold and contig levels (Figures 2C, D). It was confirmed that at least one model animal in representative species at each class taxonomic level, such as yeast, Drosophila, chicken, and frog, has an extremely high complete genome.

Since various types of repeat elements widely spread across the genome are a representative cause of difficulty in the genome assembly process, we further investigated the frequency of repeat sequences in the genome of each species to evaluate the quality of each reference genome. We hypothesized that genome repeat frequencies in each species could help assess the quality of the reference genomes; however, there was no association with various genome quality indicators (Figure 2A). We found that one of the primary reasons for this observation is that the genome size varies across species, depending on the class taxonomic level, and that genome size determines the types of repeat elements that can be found (Figure 2E). A correlation of 0.924 was observed between the length occupied by all repeat elements in the genome and the length of the genome, supporting this claim. In addition, it is further evidence that the length of the region occupied by the repeat sequence in the entire genome is mostly dependent on long repeat sequences such as LINE1 and LINE2 (Figure 2F). Although all species had a consistent linear pattern in their genome size and ratio of repeat elements, we found that species such as *Leptobrachium leishanense* had a high ratio of repeat elements to genome size (Figure 2E). However, since we cannot be sure whether these results are due to the characteristics of the genome of the species, we ultimately concluded that it is difficult to use the ratio of repeat elements as an effective measure to evaluate the quality of the genome. Additionally, we used BUSCO to compare the quality of reference assemblies of multiple species based on the orthologous genes. In result, all BUSCO completeness in 109 species had high values (97.255 in average) with no significant differences, which means that there is no value as an effective indicator for comparing multiple species with reference genomes (Supplementary Figure 3).

Demonstration of change in the mapping quality of RNA-seq data according to the completeness of the reference genome

We demonstrated whether the representative indicators used to evaluate the quality of the reference genome affect the mapping step of RNA-seq data processing. For the remaining 108 species, excluding *Salmo trutta*, for which repeat elements were not

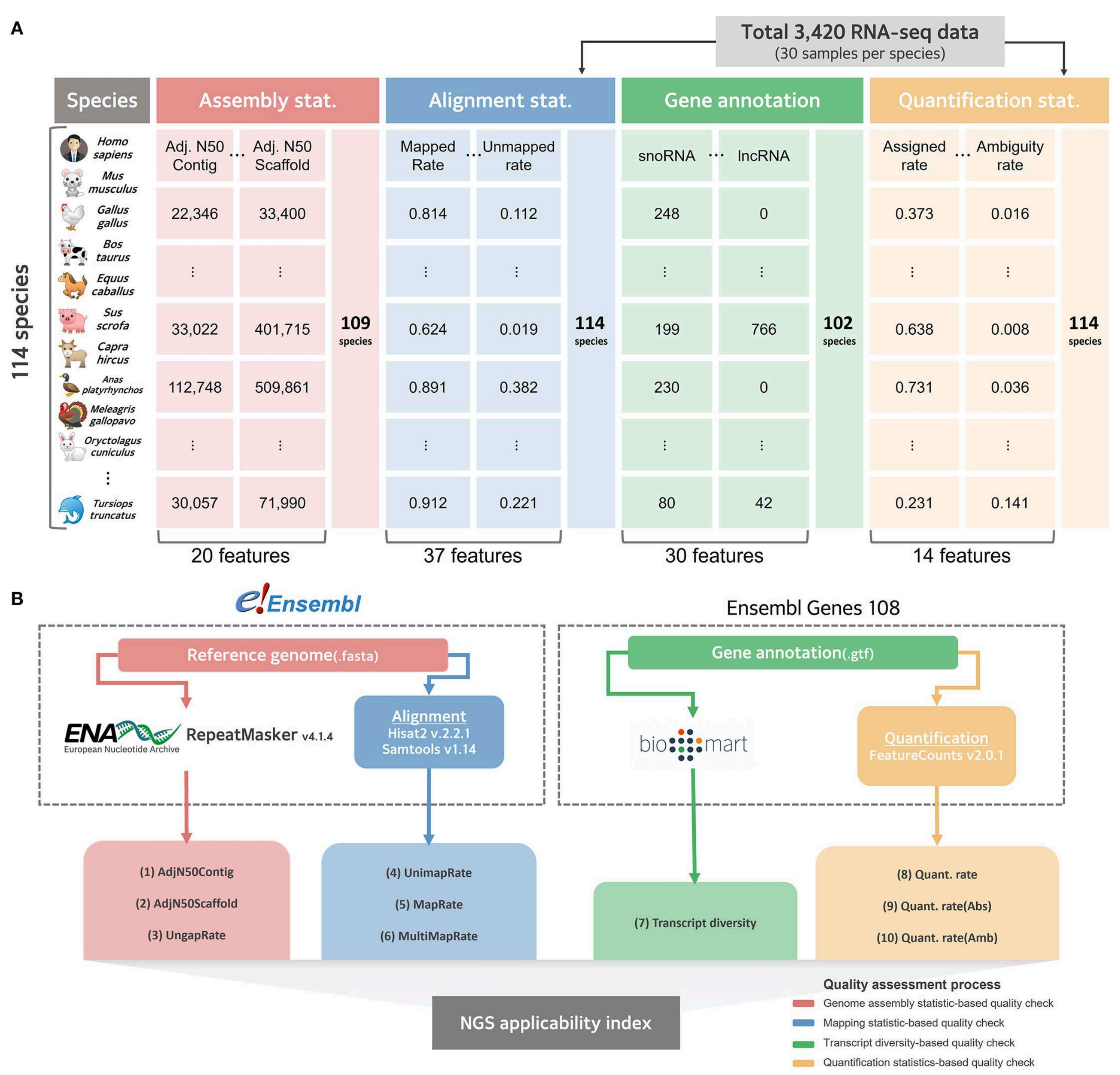


FIGURE 1
Collected data structure and schematic diagram for benchmarking comparison. **(A)** Overall structure of data collected for quality evaluation of reference genome and gene annotation for 114 species. **(B)** A systematic workflow to select effective indexes for relative quality assessment from collected data. The red line represents a pipeline that selects effective indicators from assembly statistics for relative evaluation of the reference genome. The blue line represents the process of empirically evaluating the quality of the genome by mapping the actual NGS data in the alignment step. The green line represents the process of calculating the transcript diversity index, and the yellow line represents the pipeline that empirically finds effective indicators for quality evaluation of gene annotations in the quantification process.

identified among 109 species, 3,240 RNA-seq data were mapped to their corresponding reference genome in a non-repeat masked version. Although no clear linear relationship was observed when the characteristics of different species were considered simultaneously, we found that the mapping failure rate increased, and the unique mapping and total alignment rates decreased as the number of spanned gaps increased (Figure 3A). Similarly, in another assembly contiguity index, with N50, N75, and N90 adjusted by genome size, it was demonstrated that the mapping failure rate decreased, and the mapping success rate increased when longer contig or scaffold values were observed. These results provide evidence that the quality of the mapping step is directly affected by the genome completeness.

We also demonstrated that the multiple mapping problem intensifies depending on the ratio of the repeat elements in the genome. It was demonstrated that the rate of multiple mapping reads increased (r^2:0.394) in genomes with a high frequency of repetitive sequences across all species (Figures 3A, B).

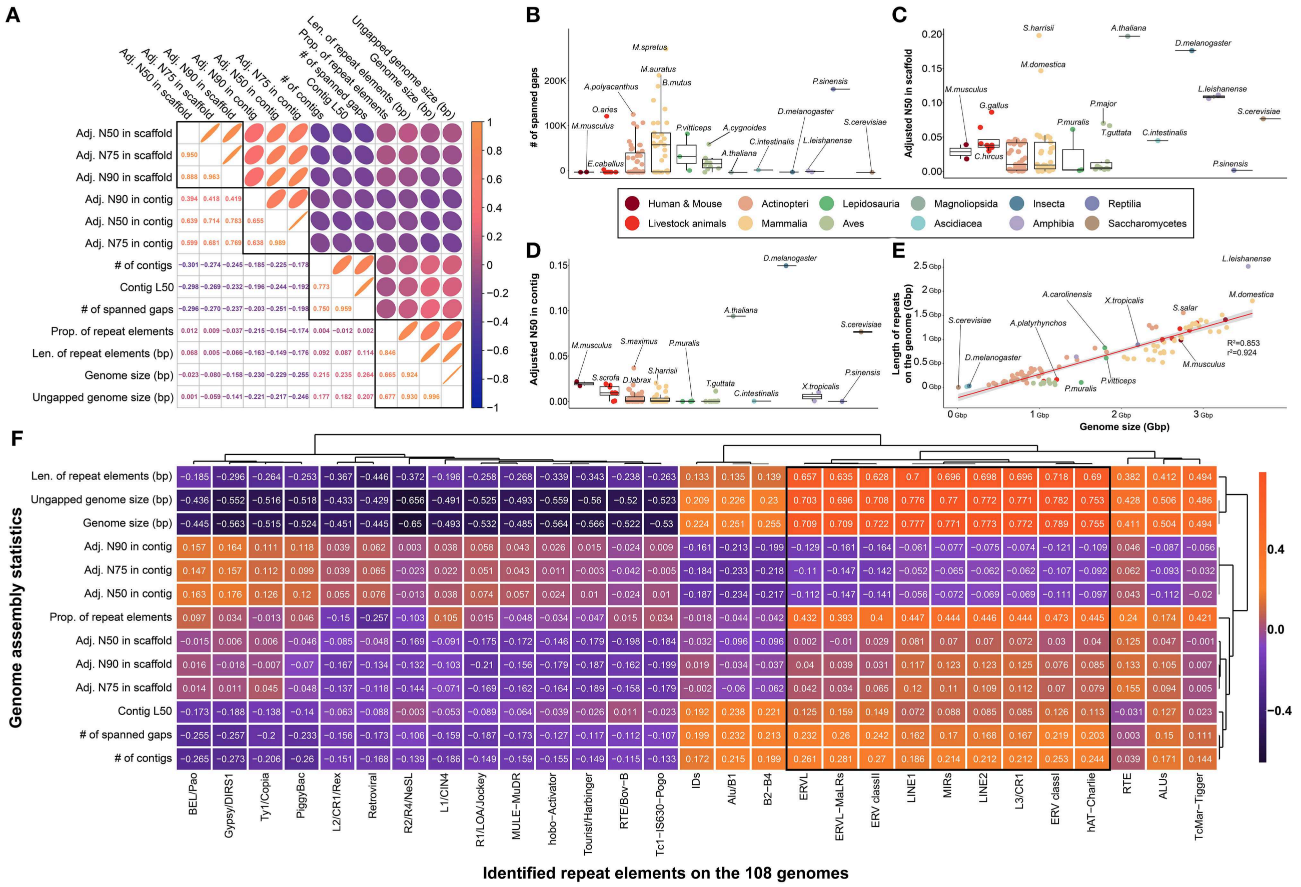


FIGURE 2
Comparison of assembly statistics for selection of effective indicators for genome quality evaluation of various species relatively. **(A)** Investigation of correlation between various assembly statistics and repeat elements that are presumed to be related to genome quality evaluation. Pearson's correlation coefficients were used to pairwise investigation. Four major types of indicators showed strong correlations. **(B)** The number of spanned gaps remaining in the genomes of 109 species. **(C)** Comparison of adjusted N50 in scaffold level by genome size. **(D)** Comparison of adjusted N50 in contig level by genome size. **(E)** Strong correlation between genome size and total length of repeat elements in each species. R^2 and r^2 represents coefficient of determination and correlation coefficient, respectively. **(F)** Correlation between the assembly statistics and the amount of various types of repeat elements found in the genome of 108 species. **(B–E)** The colors in the figure share group information separated by class taxonomic level, except for human-mouse and eight livestock animals.

This is because the genome used in this experiment was an unmasked version of the repeat elements. If the genome utilized repetitive masked versions commonly used in RNA-seq, the multiple mapping rate would not increase, but the overall mapping rate would decrease. The average multi-mapping rate in all species was 5.68%, whereas a multi-mapping rate of 22.512% was observed in *Xenopus tropicalis*. High multi-mapping rates were also observed in model organisms such as *D. melanogaster* (15.068%) and *A. thaliana* (13.034%). These results demonstrate that multi-mapping of reads intensifies according to the ratio of repeat sequences in the genome, however this could be because of the characteristics of the species, not the quality of the genome (Figure 3B).

Finally, we compared all species with MQI based on valid indicators generated in the mapping step. An average MQI of 0.829 was observed across all species, indicating that there are very few species with genomes that perform poorly enough to affect mapping in most publicly available reference genomes (Figure 3C).

Qualitative evaluation of gene annotations from 102 species through comparison of transcript diversity

Based on 30 different types of genes included in the gene annotation collected from a total of 102 species (Supplementary Table 9), we evaluated the relative level of gene annotation in various species, including livestock. We hypothesized that the gene annotations for humans and mice, which have been frequently and continuously revised through the efforts of many researchers, would be at the highest level. The fact that 24 of the 30 classification criteria of the transcript types in gene annotation were observed in human and mouse species demonstrates that this is the most subdivided gene annotation when compared to other species, as we hypothesized (Figure 4A). Therefore, it was further hypothesized that by measuring the transcript diversity of gene annotation within a specific species, it would be possible to measure the relative level of gene annotation

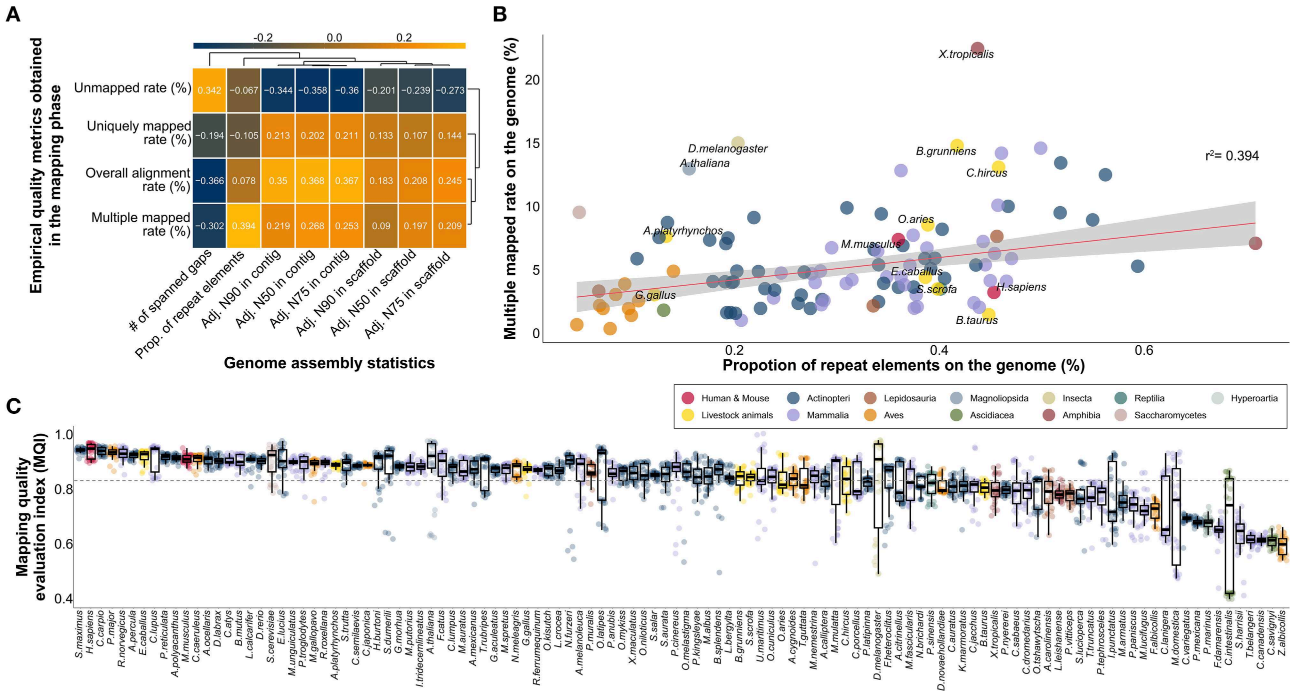


FIGURE 3
Investigation of association relationship between assembly statistics and empirical effective indicators obtained in the mapping step for genome quality evaluation. **(A)** Pairwise correlation between selected assembly statistics and empirical effective indicators obtained in the mapping step for genome quality assessment. **(B)** Linear relationship between the ratio of multiple mapped reads and proportion of repeat elements on the genome in 108 species. r^2 represents correlation coefficient, respectively. **(C)** Differences in newly proposed MQI values in 108 species. To evaluate the relative quality of the genome, valid empirical indicators were integrated and configured in the mapping stage. The horizontal line represents the average MQI value across all 108 species.

of that species compared to humans or mice, which have relatively well-organized gene annotations.

As a result of investigating gene diversity in annotations using a dimensionality reduction algorithm based on the ratio of 30 different types of genes derived from 102 gene annotations, no species has yet reached the level of human or mouse gene annotation (Figure 4B). The PC1 values obtained from dimensionality reduction analysis explained 77.92% of the total transcript diversity in gene annotations, and the strong correlation with Shannon's equitability calculated based on mouse species supports our claim (Figure 4C). We evaluated the diversity of transcripts in each of the 102 gene annotations and found the highest diversity in human (Figure 4D). Based on human's transcriptome diversity, mouse gene annotation followed with 87.996%. In the case of mammals, the average diversity of gene annotations was generally higher than that of other classes. Livestock were confirmed to have approximately 39.463% diversity compared to that of the human gene annotation. Of the eight livestock species highlighted in this study, only 12.099% of the human gene annotation complexity was annotated in the mallard duck (*Anas platyrhynchos; A. platyrhynchos*). While gene annotations with more than 50% transcript diversity were rare in other classes, relatively high gene annotation diversity levels of 48.095% and 47.999% were found in *D. melanogaster* and *Salmo salar*, respectively.

We further investigated whether the transcript diversity index was significantly affected by which of the 30 transcript types (Supplementary Table 10). It was found that lncRNA had a correlation of 0.841 with the transcript diversity obtained from the dimensionality reduction analysis. We found that lncRNAs in 37 species, including *D. melanogaster*, were not classified in the annotation (Figure 4E, Supplementary Table 9). Among the 8 livestock animals, 11,944 and 10,965 lncRNAs were annotated in *Gallus gallus* and *Sus scrofa*, respectively. In contrast, relatively low numbers of 1,480 and 786 lncRNAs were annotated in *Bos taurus* and *A. platyrhynchos*. We presumed that protein-coding genes would contribute considerably to the diversity of gene annotation, but correlation of −0.126 with transcript diversity was found (Supplementary Table 10). In addition, the average proportion of protein-coding genes was 81.69% in all 102 species (Figure 4F). These results demonstrated that when constructing gene annotations across all species, protein-coding genes are usually annotated as primary targets; thus, they did not significantly contribute to the classification of the 102 species based on the diversity of transcripts within the annotations. However, we identified relatively low proportions of protein-coding genes in model organisms such as humans (33.041%), mice (38.538%), chickens (56.487%), and *D. melanogaster* (58.365%). Concurrently, we found that various small RNAs, such as small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small Cajal body-specific RNA (scaRNA), and miRNA, also play an important role in determining the level of transcript diversity for gene annotations in 102 species (Supplementary Figure 4). This implies that as non-coding genes other than protein-coding genes are included

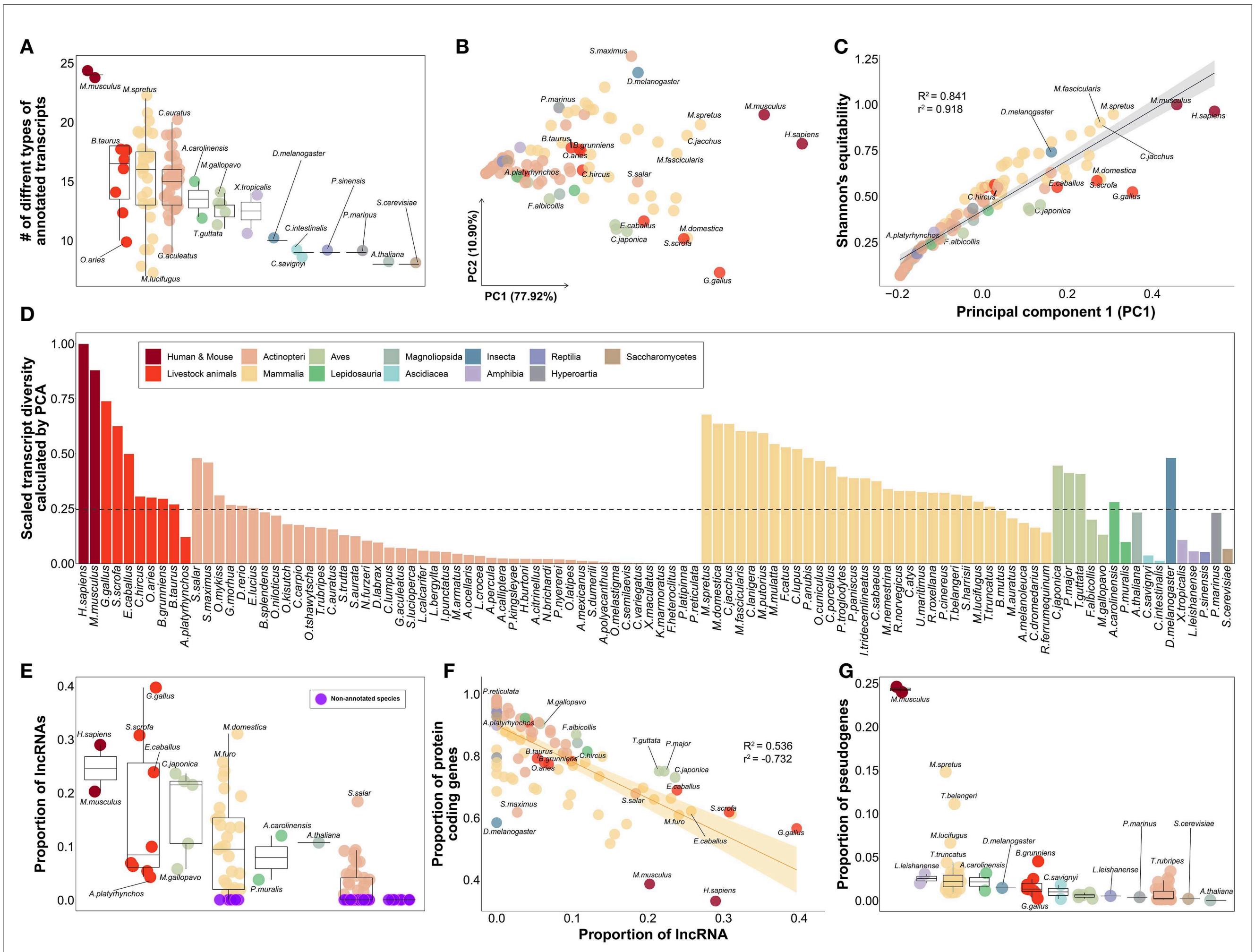


FIGURE 4

Qualitative evaluation of gene annotation based on transcript diversity. **(A)** Comparison of the number of types found in each species out of a total of 30 transcript types annotated in gene annotations of 102 species. **(B)** Dimensional reduction results for the ratio of 30 transcript types in gene annotation of each species through principal component analysis. About 77.92% of the total variance in the original data was explained by the first principal component. **(C)** Results of correlation investigation between two methods of estimating transcript diversity: Shannon's equitability and PC1 obtained through PCA. R^2 and r^2 represents coefficient of determination and correlation coefficient, respectively. **(D)** Comparison of transcript diversity index for all 102 species. **(E)** Comparison of the proportion of annotated lncRNAs in the gene annotations of each species. Purple indicates species with no lncRNA annotated at all. **(F)** Correlation between the proportion of lncRNAs and the proportion of protein-coding genes. R^2 and r^2 represents coefficient of determination and correlation coefficient, respectively. **(G)** Comparison of the proportion of annotated pseudogenes in the gene annotations of each species. **(A–E)** The colors in the figure share group information separated by class taxonomic level, except for human-mouse and eight livestock animals.

in the gene annotation, the proportion of protein-coding genes decrease, suggesting that this can be another indicator of the degree of development of gene annotation. Finally, we observed a correlation of 0.545 between transcript diversity and the ratio of pseudogenes (Supplementary Table 10). Excluding human and mouse gene annotations, the average proportion of genes classified as pseudogenes in the gene annotations of the remaining 100 species was only 1.523% (Figure 4G). In contrast, in humans and mice, a significant number of annotated genes were classified as pseudogenes, at 24.571 and 23.961%, respectively. This result indicates that the level of gene annotation is generally higher, as pseudogenes are additionally considered in gene annotation beyond the level of simple classification of protein-coding genes, lncRNAs, and some small RNAs whose functions are known or are of common interest to scientists.

Demonstration of change in mapped reads quantification performance according to the quality of gene annotation in 102 species

Quantification of reads generated from RNA-seq data is a crucial process for measuring gene expression levels and is most frequently applied to various biomedical fields. In the process of quantifying the reads mapped to the genome, we speculated that the quantification success rate would be affected by the structure and completeness of the gene annotation of various species. Based on RNA-seq data from all 102 species, we found that the proportion of annotated exon ($r^2 = 0.45$) or gene ($r^2 = 0.473$) in the genome correlated most with the proportion successfully assigned to a

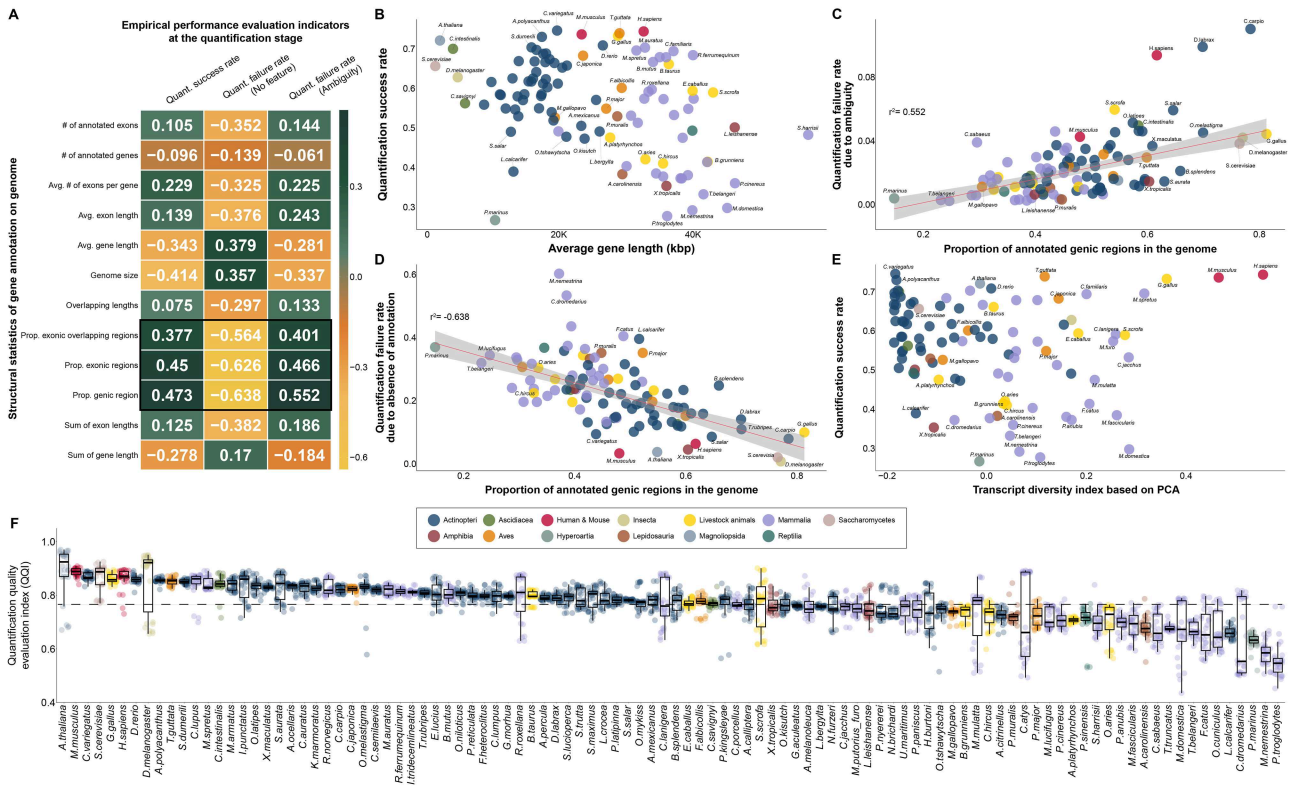


FIGURE 5
Selection of empirical effective indicators in the quantification process and investigation of correlation with complexity of transcripts in annotation for quality evaluation of gene annotation in diverse species. **(A)** Correlation between 12 characteristics of gene annotation and 3 quality evaluation indicators obtained empirically in the quantification process. **(B)** Scatter plot between average gene length and quantification success rate. **(C)** Association between the proportion of annotated genic regions in the genome and the rate of quantification failure due to ambiguity. **(D)** Correlation between the percentage of annotated genic regions in the genome and the rate of quantification failure due to the absence of annotation. **(E)** Independence between transcript diversity index, a proposed qualitative quality evaluation metric, and quantification rate, a quantitative quality evaluation index. **(F)** Differences in QQI, an empirical quality index obtained at the quantification stage, in all species. **(B–F)** The colors in the figure share group information separated by class taxonomic level, except for human-mouse and eight livestock animals.

specific gene during the quantification process (Figure 5A). We found patterns clearly differentiated by average gene length in 102 species at the class taxonomic level and identified the quality of gene annotation within each class in terms of the quantification rate for mapped reads on the genome (Figure 5B). For example, human (0.745) and mouse (0.738) gene models are of outstanding quality in mammals; however, the quantification rates were significantly low in *Macaca nemestrina* (0.293) and *Pan troglodytes* (0.279). High quantitative success rates were observed in *G. gallus* (0.734) and *A. thaliana* (0.722), which are representative model bird and plant species, respectively. However, in *Petromyzon marinus* (0.268), which represents the Hyperoartia class, it was confirmed that RNA-seq application research is not yet possible in terms of the quantification rates of mapped reads.

While genomic features were distinct for each class taxonomic level, we found a common pattern across 102 species in two representative causes of mapped reads for which quantification failed (Figures 5C, D). The first representative cause of quantification failures caused by gene annotation was ambiguity due to redundant annotations at genomic locations (Figure 5C). We demonstrated that a higher percentage of genes annotated in the genome of a particular species, led to increased ambiguity (r^2 = 0.552) in the quantification step (Figures 5A, C). Interestingly, it was also found that human and mouse gene annotations, which had a high quantification success rate, were not free from redundancy problems, suggesting that short-read-based NGS technology continue to have difficulties in accurate quantification. We further investigated the absence of gene annotation, which is another representative cause of quantification failure for mapped reads caused by gene annotation. As a result, we identified a common pattern in which higher frequency of genes annotated in the genome, led to the lower quantification failure rate (r^2: −0.638) due to the absence of annotation (Figure 5D). We demonstrated that in most model organisms, including humans (0.065) and mice (0.035), the rate of quantification failure caused by the absence of gene annotation was relatively low compared to that in other species. We also demonstrated that these two representative quantification errors (Figures 5C, D), caused by the characteristics of gene annotation, were opposed to each other in 102 species through actual RNA-seq data. For example, human and mouse annotations include annotations for many genes compared to other species, reducing errors due to the absence of annotations; however, errors due to redundancy of annotations are relatively high. In this regard, we additionally investigated the association

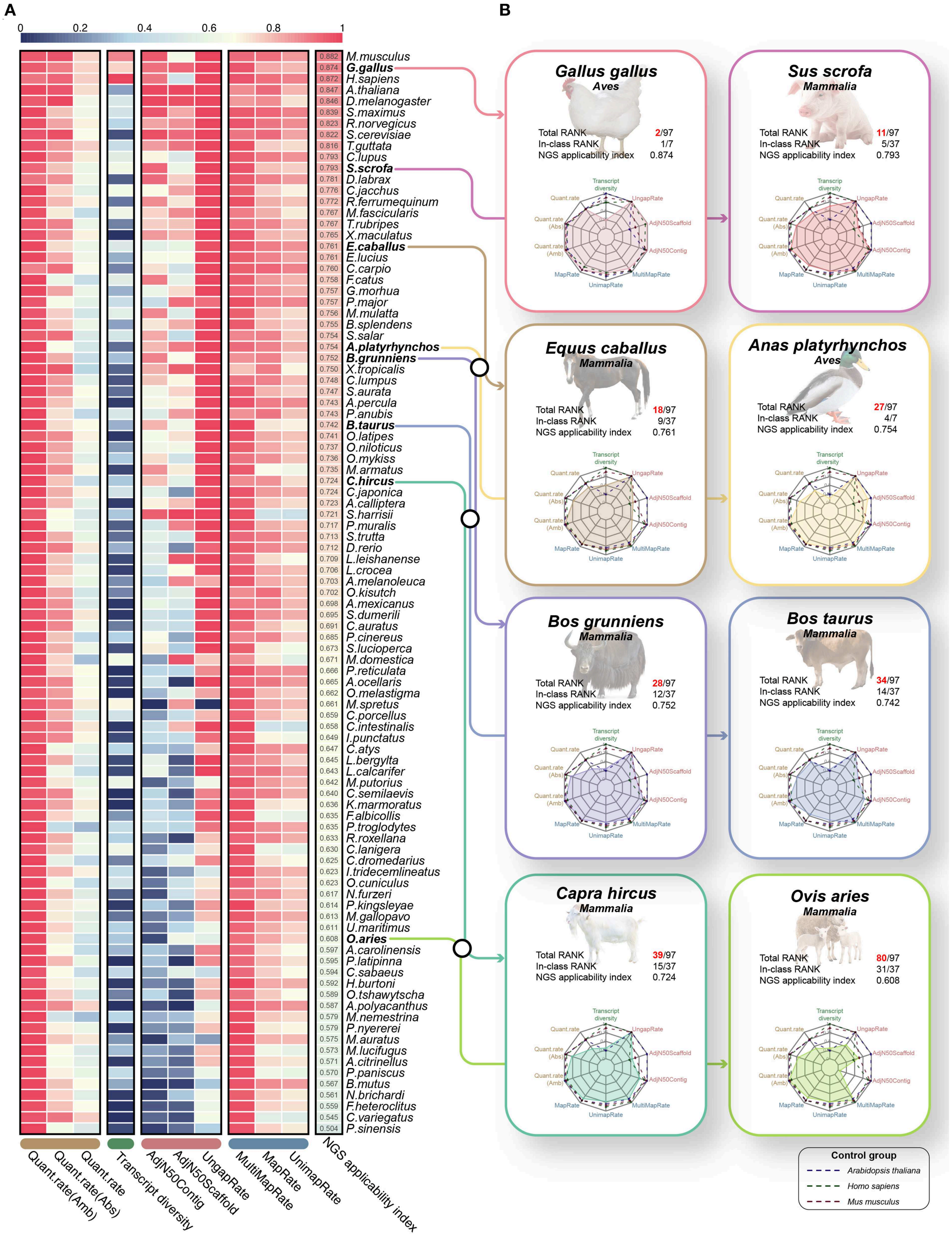


FIGURE 6
Quality evaluation results of 97 species through the proposed NGS applicability index based on the 10 quality evaluation indicators verified through this study. **(A)** Heatmap for a total of 10 quality evaluation indicators selected through this study. The heatmap includes three assembly evaluation indicators and three performance indicators derived from the mapping process, which can relatively evaluate the quality of genomes. In addition, transcript diversity and three performance indicators derived from the quantification process are included to relatively evaluate the gene models. Finally, all 97 species were sorted in descending order through the NGS applicability index, which is the result of the weighted sum of these 10 quality evaluation indicators. All values have a scale of 0.0 to 1.0, and the closer to 1, the higher the quality. **(B)** Results of benchmarking quality evaluation of reference genome and gene annotation for 8 livestock animals.

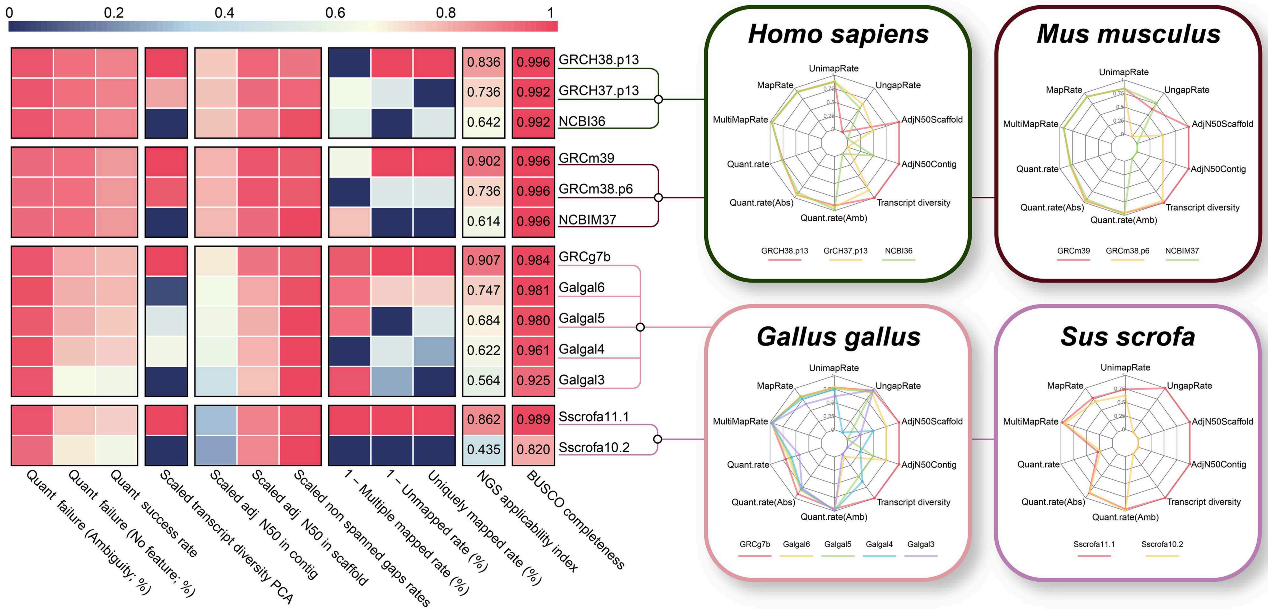


FIGURE 7
Technical validation of the NGS applicability index based on the different genome builds. Comparison between NGS applicability index and BUSCO completeness for different version of reference genome and gene annotation. Polygonal charts represent values for each of the 10 effective indicators that make up the NGS applicability index. The larger the polygon area represent the higher the NGS applicability index.

between the diversity of annotated transcript types and the success rate of quantification, but no association was observed (Figure 5E). This result demonstrated that the transcript diversity index does not affect the quantification success rate index, as it does not affect the exon or gene structure in gene annotation. In addition, the transcript diversity index has been demonstrated to be another independent index that can evaluate gene annotation qualitatively in a different direction than the quantification success rate index.

We finally compared a QQI for 102 species based on the quantification success rate and two quantification failure rates, which are determined by the quality of gene annotation (Figure 5F). As a result, it was found that the average QQI was high in the order of *A. thaliana* (0.89), mouse (0.887), *C.variegatus* (0.871), *S.cerevisiae* (0.866) and chicken (0.863). This result demonstrates that most model organisms whose gene annotations have been frequently updated are of markedly high quality compared to other species through the quantification process with real 3,060 RNA-seq data from 102 species. In contrast, this suggests that there are still practical problems with accurate quantification due to quality problems of gene annotation in species belonging to Mammalia, such as *Camelus dromedarius* (0.644), *Macaca nemestrina* (0.585) and *Pan troglodytes* (0.557).

Application and validation of NGS applicability index

Finally, we proposed the NGS applicability index by integrating 10 validated effective indicators that can evaluate the reference genome and gene annotation (Figure 6A, Supplementary Table 11). As a result, mice (0.882), chickens (0.874), humans (0.872) and *Arabidopsis* (0.847) species were observed in the order of highest scores (Figure 6A), which revealed that the NGS applicability index is valid for relative quality assessment in diverse species. We expected that through this NGS applicability index, we could evaluate the boundaries of NGS application research and the direction of development to improve the quality of the genome and gene annotation for a specific species. For example, although *Arabidopsis* and turbot showed extremely high NGS applicability indices, transcript diversity was 0.233 and 0.46, respectively, compared to other high-ranking species. From this, there is no technical problem in performing applied NGS technologies, such as whole genome resequencing or RNA-seq, but it is not possible to study various types of transcripts, including lncRNAs and various small ncRNAs. Simultaneously, it can be understood that these species will improve the direction of increasing the transcript diversity of gene annotations, such as diverse ncRNAs. An integrated quality index of 0.751 on average was observed in all eight livestock animals, it has not yet reached the level of other model animals except for chickens, suggesting that it has stable quality compared to other species (Figure 6B). Because relatively low quantification success rates are observed in goats, yaks, and sheep, gene annotation must be improved soon.

Generally, when an assembly build is upgraded, a significant increase in the quality of the reference genome and/or gene annotation is expected. Taking this into account, we additionally compared different assembly builds from four species with a high NGS applicability index (human, mouse, chicken, and pig) to verify the validity of the proposed NGS applicability index. As expected, as the genome build increased in all four species, the NGS applicability index improved significantly (Figure 7), which is direct evidence supporting the validity of our proposed quality indicator. The BUSCO completeness, a representative methodology for evaluating genome assembly, also showed a tendency to increase as the assembly build improved, but it was observed that the difference was relatively insignificant. In particular, the NGS applicability index showed a clear increase in the order of 0.624, 0.745, and 0.912 for the mouse, but the BUSCO Completeness was the same at 0.996. This result is direct evidence that the NGS applicability index can show higher quality assessment discernment by simultaneously considering more diverse aspects than the BUSCO method, which focuses only on the completeness of genome assembly (Supplementary Figure 5).

Discussion

To date, various studies have been conducted to compare and evaluate the quality of genomes and gene annotations; however, most have been used to compare evolutionarily close species (10, 43) or assembly methods (44, 45). Since most studies have aimed at comparing adjacent minority species, the quality evaluation indicators that have been used are limited, and discussion on the methodology to compare genomes and gene annotations of multiple species is lacking. However, reference genomes and gene annotations are essential data for various NGS application technologies, including RNA-seq data, and have been known to directly affect the performance of essential steps, such as alignments and quantification processes (28, 31). While the application of NGS technology in various species is becoming increasingly common, the quality of these key data can influence the accuracy of the research outcome itself; therefore, it must be evaluated. In this study, genomes and gene annotations of 114 species, including eight livestock species, were obtained from the Ensembl database, and 3,420 RNA-Seq data were collected to attempt diversified quality evaluation in various species (Figure 1). We conducted research to find novel effective indicators for quality assessment, and to select effective indicators among existing quality assessment indexes that can objectively evaluate the genome and gene annotation of a specific species.

Among the indicators generated in the *de novo* assembly process, which is used for quality evaluation of reference genomes, the validity of the N50 values of contig and scaffold levels was first examined (Figure 2). This N50 value, called the contiguity index, refers to the length at which contigs or scaffolds are sorted in length order and reach 50% of the target length of the complete assembly (20). However, this value fluctuates depending on the final target length; therefore, it is not suitable for comparing multiple species with different genome lengths (46). Therefore, in most studies using the N50 index, the genome size of the target species is usually unknown, and has been used to compare the quality of the genome assembly by estimation based on the genome size of evolutionarily close species (21, 22). Because the genome sizes were fixed for the purpose of our study, we converted the N50 value to an effective index that can be compared between multiple species by correcting it with the genome size of the species. As a result, we identified an association with the quality index that directly indicates the quality of the reference genome, such as the number of gaps in the genome (Figure 2A). This gap is the primary target in all reference genome construction studies, and various attempts have been made to minimize it (23–25). We additionally assumed that the repeat elements spread on the genome could be considered as quality indicators; however, the distribution of repeat elements is determined by the characteristics of the species (47) and thus could not be employed as another objective quality indicator (Figures 2E, F). Going one step further, we demonstrated that the three selected genome quality evaluation indicators directly affected the mapping stage of the actual NGS application (Figure 3A). In addition, the genome quality of various species can be evaluated from another perspective through the MQI score, which was created by composing indicators empirically obtained in the mapping step, such as alignment success and failure rate, and failure rate due to multiple mapping (Figure 3C). In conclusion, we selected adjusted N50 values in contig and scaffold levels, number of spanned gaps, and MQI, which are effective indicators for evaluating the quality of reference genomes of various species.

Multiple methods exist for measuring the quality of a reference genome, but the only way to measure the completeness of annotated transcripts in the genome is to compare them with the annotations of evolutionarily similar species (18, 48). In other words, because there is no objective indicator for the quality evaluation of gene annotation, it was not possible to evaluate the quality of various species. In this context, we proposed a novel metric, transcript diversity, to evaluate the completeness of gene annotation in various species (Figure 4). We calculated the diversity of this transcript under the assumption that gene models frequently developed by multiple scientists, such as humans or mice, would eventually be of the highest quality. As evidence for this, we demonstrated that gene annotations in humans and mice are fine-grained for lncRNAs (Figure 4E), various small RNAs, and pseudogenes (Figure 4G). In the past, the elucidation of protein-coding genes has been a major goal, even in representative gene models, including humans and mice (49). However, as it was revealed that non-coding genes such as various types of lncRNA (50), snRNA (51), snoRNA (52), scaRNA (53), and miRNA (54) are also involved in various functions in living organisms, more diverse transcript types have been included in gene annotation of human and mouse. Considering the developmental history of this representative gene model, we believe that our newly proposed transcript diversity has sufficient value as a new index to measure the quality of gene annotation. In addition, we showed that transcript diversity, a qualitative quality indicator, was independent of QQI, a quantitative quality indicator of gene annotation (Figure 5A). Like the MQI, an empirical quantitative index that can evaluate the quality of the genome in the mapping stage, we proposed QQI as a novel indicator, which can evaluate the quality of gene annotation in the quantification process. We demonstrated that the success rate of quantification of mapped reads and both failure rates depended on the complexity of each gene annotation (Figure 5). This is strong evidence to show that the QQI, which is the sum of these three empirical indicators, is also an indicator that can evaluate gene annotation from a different perspective than the transcript diversity index (Figure 5F). In conclusion, we present a novel transcript diversity index, a qualitative index that can evaluate the gene annotations of various species, and the QQI, a quantitative index that can be empirically evaluated. We also demonstrated that they can be used to evaluate the quality of gene annotation in diverse species.

In this study, we attempted a novel approach to compare the quality of reference genomes and gene annotations of multiple species; however, there were limitations. First, we limited the number of species to those from which could collect more than 30 samples of RNA-seq data from species listed in the Ensembl database. If additional species are considered , there is a possibility that the evaluation of the middle and lower ranks may change. Second, although quality control was performed as best as possible for the 30 RNA-seq data samples collected for each species, the data contained random errors, as experimentally identical tissues and environmental conditions were not controlled across all species. This factor can affect the empirical quality metrics. Third, only an intuitive scaling method incorporating 10 quality evaluation indicators was applied in this study. We believe that a methodology that can efficiently integrate heterogeneous indicators derived from these diverse species will be elucidated in near future. Lastly, we considered only those quality evaluation indicators that could be obtained from available data; information that was not publicly available, such as the mis-assembly rate or assembly depth coverage, could not be considered. However, because the relative methodology proposed in this study is a framework, these practical issues are expected to be automatically resolved as reference genomes and gene annotations for various organisms are revealed. Concurrently, the relative index will become more accurate.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SPar, JiL, and MS designed the study and wrote the manuscript. SPar, JiL, JK, DK, and MS analyzed the data. SPar and JiL collected sequencing data. All authors reviewed and edited the manuscript, contributed to the article, and approved the submitted version.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (Nos. 2021R1C1C1008199 and 2021R1A5A8032895).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2023.1128570/full#supplementary-material

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* (2008) 24:133–41. doi: 10.1016/j.tig.2007.12.007

2. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet.* (2014) 30:418–26. doi: 10.1016/j.tig.2014.07.001

3. Moss DJH, Pardiñas AF, Langbehn D, Lo K, Leavitt BR, Roos R, et al. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.* (2017) 16:701–11. doi: 10.1016/S1474-4422(17)30161-8

4. Bien SA, Su Y-R, Conti DV, Harrison TA, Qu C, Guo X, et al. Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Hum Genet.* (2019) 138:307–26.

5. Wang J, Dean DC, Hornicek FJ, Shi H, Duan Z. RNA sequencing (RNA-Seq) and its application in ovarian cancer. *Gynecol Oncol.* (2019) 152:194–201. doi: 10.1016/j.ygyno.2018.10.002

6. Lezmi E, Benvenisty N. Identification of cancer-related mutations in human pluripotent stem cells using RNA-seq analysis. *Nat Protoc.* (2021) 16:4522–37. doi: 10.1038/s41596-021-00591-5

7. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol.* (2018) 18:35–45. doi: 10.1038/nri.2017.76

8. Liang Q, Dharmat R, Owen L, Shakoor A, Li Y, Kim S, et al. Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling. *Nat Commun.* (2019) 10:1–12. doi: 10.1038/s41467-019-12917-9

9. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil QJGB. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* (2020) 21:1–16. doi: 10.1186/s13059-020-1935-5

10. Florea L, Souvorov A, Kalbfleisch TS, Salzberg S. Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. *PLoS ONE.* (2011) 6:e21400. doi: 10.1371/journal.pone.0021400

11. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. *Nucleic Acids Res.* (2023) 51:D933–41. doi: 10.1093/nar/gkac958

12. Paez, S., Kraus, R. H., Shapiro, B., Gilbert, M. T. P., Jarvis, E. D., and Vertebrate Genomes Project Conservation Group. (2022). Reference genomes for conservation. *Science*. 377, 364–366. doi: 10.1126/science.abm8127

13. Sequencing and Nature. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature.* (2005) 437:69–87. doi: 10.1038/nature04072

14. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature.* (2013) 496:498–503. doi: 10.1038/nature12111

15. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics.* (2017) 33:574–6. doi: 10.1093/bioinformatics/btw663

16. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* (2020) 21:1–27. doi: 10.1186/s13059-020-02134-9

17. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* (2021) 22:1–21. doi: 10.1186/s13059-021-02527-4

18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* (2015) 31:3210–2. doi: 10.1093/bioinformatics/btv351

19. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *In: Gene prediction.* New York, NY: Springer. (2019) p. 227–45. doi: 10.1007/978-1-4939-9173-0_14

20. Mäkinen V, Salmela L, Ylinen J. Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics.* (2012) 13:1–5. doi: 10.1186/1471-2105-13-255

21. Williams, J. L., Iamartino, D., Pruitt, K. D., Sonstegard, T., Smith, T. P., Low, W. Y., et al. (2017). Genome assembly and transcriptome resource for river buffalo, *Bubalus bubalis* (2 n= 50). *Gigascience.* 6, gix088. doi: 10.1093/gigascience/gix088

22. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants.* (2018) 4:879–87. doi: 10.1038/s41477-018-0289-4

23. Marti-Renom MA, Mirny L. Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput Biol.* (2011) 7:e1002125. doi: 10.1371/journal.pcbi.1002125

24. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol.* (2012) 13:1–9. doi: 10.1186/gb-2012-13-6-r56

25. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience.* (2012) 1:18. doi: 10.1186/2047-217X-1-18

26. Shumate A, Salzberg SL. Liftoff: an accurate gene annotation mapping tool. *Bioinformatics.* (2020) 37:1639–43. doi: 10.1093/bioinformatics/btaa1016

27. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics.* (2009) 25:3207–12. doi: 10.1093/bioinformatics/btp579

28. Saha A, Battle A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res 7.* (2018) 1860. doi: 10.12688/f1000research.17145.1

29. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, Mcpherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* (2016) 17:1–19. doi: 10.1186/s13059-016-0881-8

30. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature.* (2010) 464:768–72. doi: 10.1038/nature08872

31. Robert C, Watson MJ. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* (2015) 16:1–16. doi: 10.1186/s13059-015-0734-x

32. Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford).* 2011:bar030. doi: 10.1093/database/bar030

33. Smit A, Hubley R, Green P. *RepeatMasker Open-4.0. 2013–2015.* (2015). Available online at: http://www.repeatmasker.org

34. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit A. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* (2021) 12:1–14. doi: 10.1186/s13100-020-00230-y

35. Benson GJ. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* (1999) 27:573–80. doi: 10.1093/nar/27.2.573

36. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* (2010) 39: D19–D21. doi: 10.1093/nar/gkq1019

37. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. In: *Babraham Bioinformatics.* Cambridge, United Kingdom: Babraham Institute.

38. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnol.* (2019) 37:907–15. doi: 10.1038/s41587-019-0201-4

39. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* (2014) 30:923–30. doi: 10.1201/b16589

40. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* (1948) 27:379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

41. Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology.* (1973) 54:427–32. doi: 10.2307/1934352

42. Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, et al. (2022). The European nucleotide archive in 2022. *Nucleic Acids Res.* (2022) 51:D121–D125. doi: 10.1093/nar/gkac1051

43. Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, et al. The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun.* (2014) 5:1–9. doi: 10.1038/ncomms4833

44. Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* (2011) 21:2224–41. doi: 10.1101/gr.126599.111

45. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience.* (2013) 2:10. doi: 10.1186/2047-217X-2-10

46. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* (2010) 95:315–27. doi: 10.1016/j.ygeno.2010.03.001

47. Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay TJ. Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *J Evol Biol.* (2022) 36:321–36. doi: 10.1111/JEB.14106/v2/response1

48. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* (2009) 37:289–97. doi: 10.1093/nar/gkn916

49. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet.* (2012) 13:329–42. doi: 10.1038/nrg3174

50. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* (2008) 18:1433–45. doi: 10.1101/gr.078378.108

51. Cheng Z, Sun Y, Niu X, Shang Y, Ruan J, Chen Z, et al. Gene expression profiling reveals U1 snRNA regulates cancer gene expression. *Oncotarget.* (2017) 8:112867. doi: 10.18632/oncotarget.22842

52. Williams GT, Farzaneh F. Are snoRNAs and snoRNA host genes new players in cancer? *Nat Rev Cancer.* (2012) 12:84–8. doi: 10.1038/nrc3195

53. Ronchetti D, Mosca L, Cutrona G, Tuana G, Gentile M, Fabris S, et al. Small nucleolar RNAs as new biomarkers in chronic lymphocytic leukemia. *BMC Med Genomics.* (2013) 6:1–11. doi: 10.1186/1755-8794-6-27

54. Wang, L., Sinnott-Armstrong, N., Wagschal, A., Wark, A. R., Camporez, J.-P., Perry, R. J., et al. (2020). A microRNA linking human positive selection and metabolic disorders. *Cell.* 183, 684–701. e614. doi: 10.1016/j.cell.2020.09.017

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 12 May 2023
DOI 10.3389/fvets.2023.1160486

EDITED BY
Rajesh Kumar Pathak,
Chung-Ang University, Republic of Korea

REVIEWED BY
Ravendra Chauhan,
Rhodes University, South Africa
Divya Mishra,
University of Allahabad, India
Akshay Singh,
Indian Council of Agricultural Research (ICAR), India

*CORRESPONDENCE
Dwijesh Chandra Mishra
✉ dwij.mishra@gmail.com

†These authors have contributed equally to this work

RECEIVED 07 February 2023
ACCEPTED 11 April 2023
PUBLISHED 12 May 2023

CITATION
Mishra DC, Bhati J, Yadav S, Avashthi H, Sikka P, Jerome A, Balhara AK, Singh I, Rai A and Chaturvedi KK (2023) Comparative expression analysis of water buffalo (*Bubalus bubalis*) to identify genes associated with economically important traits. *Front. Vet. Sci.* 10:1160486. doi: 10.3389/fvets.2023.1160486

Comparative expression analysis of water buffalo (*Bubalus bubalis*) to identify genes associated with economically important traits

Dwijesh Chandra Mishra[1*†], Jyotika Bhati[1†], Sunita Yadav[1], Himanshu Avashthi[1], Poonam Sikka[2], Andonissamy Jerome[2], Ashok Kumar Balhara[2], Inderjeet Singh[2], Anil Rai[1] and Krishna Kumar Chaturvedi[1]

[1]ICAR-Indian Agricultural Statistics Research Institute, Indian Council of Agricultural Research (ICAR), PUSA, New Delhi, India, [2]ICAR-Central Institute for Research on Buffaloes, Indian Council of Agricultural Research (ICAR), Hisar, India

The milk, meat, skins, and draft power of domestic water buffalo (*Bubalus bubalis*) provide substantial contributions to the global agricultural economy. The world's water buffalo population is primarily found in Asia, and the buffalo supports more people per capita than any other livestock species. For evaluating the workflow, output rate, and completeness of transcriptome assemblies within and between reference-free (RF) *de novo* transcriptome and reference-based (RB) datasets, abundant bioinformatics studies have been carried out to date. However, comprehensive documentation of the degree of consistency and variability of the data produced by comparing gene expression levels using these two separate techniques is lacking. In the present study, we assessed the variations in the number of differentially expressed genes (DEGs) attained with RF and RB approaches. In light of this, we conducted a study to identify, annotate, and analyze the genes associated with four economically important traits of buffalo, *viz*., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency. A total of 14,201 and 279 DEGs were identified in RF and RB assemblies. Gene ontology (GO) terms associated with the identified genes were allocated to traits under study. Identified genes improve the knowledge of the underlying mechanism of trait expression in water buffalo which may support improved breeding plans for higher productivity. The empirical findings of this study using RNA-seq data-based assembly may improve the understanding of genetic diversity in relation to buffalo productivity and provide important contributions to answer biological issues regarding the transcriptome of non-model organisms.

KEYWORDS

water buffalo, transcriptome, annotation, GO terms, SSRs

Introduction

The domestic water buffalo (*Bubalus bubalis*) marks a key impact on the global agricultural economy through milk, meat, and draft power. The world's water buffalo population is largely found in Asia, and most people consider it the most promising livestock species for their livelihood (1, 2). Asia accounts for 97% of the total buffalo production with the largest population in India (>100 million) (2). More than half of the milk produced in

India comes from buffaloes, which also produce milk with higher levels of fat, particularly saturated fatty acids, than cattle (2). Buffaloes are resilient to the harsher environment and resistant to several bovine tropical diseases (1), thus may have better feed convergence while surviving on poor-quality roughage than cattle. Recent studies cataloging differentially expressed genes (DEGs) and variants (3, 4) with respect to important performance traits in water buffalo corroborate with the functional genetic diversity in this species.

Milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency traits define the overall productivity of buffaloes. Buffalo milk has the significance of having higher concentrations of fat, lactose, protein, ash, calcium, and vitamins A and C while having lower concentrations of cholesterol and the blue-green pigment (biliverdin) (5). Additionally, buffalo milk has bioactive pentasaccharides and gangliosides, which are absent in cow milk (6). Therefore, the study of genes related to milk volume is very important. The age at first calving can be used to determine a buffalo's fertility and productivity. Buffalo productivity is affected by delayed puberty onset and inadequate consecutive estrus detection (7). Reproductive efficiency is the primary factor affecting the productivity of buffaloes, which comprise early age at first calving (AFC) and optimum service period between the calvings (post-partum cyclicity) throughout the reproductive span in life. Thus, the identification of genes and the variants associated with these traits may support selective breeding for genetic improvement. Improved immunity is equally significant to propagate uterine cleansing to facilitate an early resumption of ovarian cyclicity (8, 9). Feed conversion efficiency (FCE) is defined as a dry matter intake (DMI) per unit body weight (g/day) gain determined as residual feed intake (RFI). It represents the difference in actual and predicted DMI of each individual heifer (10). Feed conversion efficiency is a heritable trait governed by common biomolecules as growth hormones are associated with milk volume, age at first calving, and post-partum cyclicity (11, 12).

Several studies have confirmed the discovery of differentially expressed as well as novel genes in mammals such as humans, buffaloes, sheep, goats, and pigs (3, 13–17). The genetic link and diversity among various buffalo breeds have primarily been studied using restriction fragment length polymorphism (RFLP) (18), random amplified polymorphic DNA (RAPD) (19), single nucleotide polymorphism (SNP) (4, 20), and simple sequence repeat (SSR) (21) markers. SSR markers have proven to be an incredibly powerful tool for researching genetic divergence and/or genetic resource conservation (22). Identification and characterization of genome-wide DEGs related to reproduction and production traits can be widely used for selective breeding, which may enhance productivity in buffaloes (23, 24).

Considering this, an attempt was made to identify the variants related to important traits, i.e., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency, using transcriptomic data to improve breeding plans in water buffaloes. DEGs were identified, characterized, and annotated, in order to accelerate performance in buffaloes through molecular breeding. This is a unique study identifying the functional classifications of genes, variants, and SSRs related to desired traits in *B. bubalis*.

Materials and methods

Milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency were the four different traits for which datasets were collected. Four samples were selected for each trait (two each of low and high expression). The complete workflow of the study is presented in Figure 1.

Ethics statement

Animals ($n = 16$) were selected as per the referred design for selective genotyping of buffaloes, based on performance phenotype recorded at ICAR-Central Institute for Research on Buffaloes (ICAR-CIRB), Hisar, Haryana, India. Genotype data were generated for selected genotypes through outsourced services hired by the institute. Animals were maintained under farm management at the institute, and the experiment design was approved by the Institute Animal Ethics Committee (IAEC) with ethics approval number—406/GO/RBI/L/01/CPCSEA.

Animals and tissue collection

Whole blood tissues of individual animals were selected from unrelated pedigrees having extreme performance levels for complex traits as follows: milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency. Each trait had four samples comprising two each of low and high expressions.

RNA extraction, library preparation, and sequencing

For the transcriptome analyses of expression patterns in low- and high-performing Murrah buffaloes, cDNA was generated using a routine RNA library preparation HiSeq protocol developed by Illumina Technologies (San Diego, CA), using 1 μg of total RNA as input. Using the High-Capacity cDNA Reverse Transcription kit (Life Technologies, Frederick, Maryland, USA), mRNA was first isolated from total RNA by performing a polyA selection step, followed by the construction of paired-end sequencing libraries with an insert size of ∼300 bp. In brief, polyA selected RNA was cleaved as per Illumina protocol, and the cleaved fragments were used to generate first-strand cDNA using Super Script II reverse transcriptase and random hexamers. Subsequently, the second strand cDNA was synthesized with RNaseH and DNA polymerase enzyme, followed by adapter ligation and end-repair steps. The resulting products were amplified via PCR, and cDNA libraries were then purified and validated using the Agilent 2200 Tape Station system (Agilent Technologies Brasil Ltda, São Paulo, SP, Brazil). Paired-end sequencing was performed using the Illumina HiScanSQ platform. Samples were multiplexed with unique hexamer barcodes and run on multiple lanes to obtain 2 × 100 bp reads. Paired-end FASTQ files were subjected to standard quality control based on Phred scores of >20, using the NGSQC Tool kit v2.2 (25) to obtain high-quality (HQ) filtered reads.

Transcriptome assembly and optimization

Raw reads from the four sets generated from the animal samples using Illumina HiSeq were filtered to generate clean data to remove adaptor sequences, reads with ambiguous sequences "*N*," low-quality reads, and reads that were mostly repeated bases, such as polyT tracts using Trimmomatic 0.39 (26). The trimmed reads are evaluated with FastQC (27), a Java-based, quality control tool for high-throughput sequence data.

After obtaining clean reads and quality checks, RF transcriptome assembly was conducted with Trinity software v2.8.6 with default parameters (28, 29). Only assembled transcripts with lengths of >300 bp were included in further analysis.

Simultaneously, the raw reads were mapped to the *B. bubalis* genome (UOA_WB_1 accessed from https://www.ncbi.nlm.nih.gov/assembly/GCF_003121395.1/) with TopHat2 v2.0.13 (30), using Bowtie2 v2.2.6 (31) as the underlying aligner. Reads aligning to the UOA_WB_1 build were quantified, which disregarded any read/read pair that aligned to more than one location or more than one gene at a single location.

Differential expression analysis

For the RF assembly results, differential transcript expression for different datasets was calculated using an exact test in the Bioconductor R package (32) edgeR [Empirical Analysis of Digital Gene Expression Data in R (33)]. We used RSEM [RNA-Seq by Expectation-Maximization (34)] to generate read counts for the optimized assembled transcriptome to input into edgeR. EdgeR normalizes raw input data using a trimmed mean of M-values (TMM), and transcripts with artificially low counts (<1 count across all samples) after normalization were excluded before differential expression analysis was completed. The transcript level was quantified in terms of Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Differential expression (DE) was detected using the edgeR Bioconductor package with a $\log_2$ fold change threshold of 2.

Differential expression analysis for the *B. bubalis* RB assembly was conducted using the Cufflink analysis tool between different samples of the same traits in pairs (high and low yielding). DE genes with $\log_2$ fold change of ≥2.0, an adjusted *p*-value (padj) of < 0.05, and an adjusted FDR of <0.05 were subjected to further analyses.

Functional annotation of genes

Functional analysis of the DEGs was performed using Blast2GO v 2.5 (35). Blast2GO is a gene ontology-based annotation tool and found to be effective in the functional characterization of sequence data. The DEGs homologous with annotated proteins in the nr database were selected for functional characterization based on the maximum *E*-value (1E-3) and the minimum alignment size (HSP length 33) using BLASTX. The DEG sequences were then categorized according to the GO vocabularies into three categories, i.e., molecular function, biological process, and cellular component. The distribution of GO terms was analyzed at level 2 of the Directed Acyclic Graphs. Annotated DEGs were analyzed for pathway identification using KEGG.

SSR mining

Simple sequence repeats were identified using the MISA tool from the DEGs. The chromosome-wise distribution of DEGs, SSRs, and SNPs [extracted from DDRAD sequence data (4)] was graphically mapped using the CIRCOS (version 0.69) visualization tool.

Results

Analysis of our data with both cattle and water buffalo reference assemblies gave varied results for differential expression and annotation among different traits, viz., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency.

Assembly benchmarking

A total of 857 million raw reads were generated (428,671,371 paired-end reads) by Illumina sequencing of the 16 *B. bubalis* samples for four traits, viz., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency, with an average of ∼26.7 million reads per sample. From these, ∼773 million reads (90.7%) were attained after removing the adapters and trimming for quality. These post-cleaning reads passed the minimum quality standards of FastQC.

After read filtering, clean reads were assembled into 488,811, 86,054, 451,596, and 451,596 contigs, reaching a total length of 482,785,524 bp, 81,971,386 bp, 431,765,546 bp, and 529,684,800 bp for "milk volume," "age at first calving," "post-partum cyclicity," and "feed conversion efficiency" traits, respectively. The average length of assembled contigs was 406, 383, 390, and 405 bp and N50 of 1,606, 1,728, 1,588, and 1,588 bp for "milk volume," "age at first calving," "post-partum cyclicity," and "feed conversion efficiency" traits, respectively.

During the course of the abovementioned analysis, the *B. bubalis* genome was sequenced in 2019, and we proceeded with the RB assembly to compare the two assembly results. To evaluate the assembly quality, we mapped the Illumina clean reads on the water buffalo reference genome (UOA_WB_1, Accession GCA_003121395.1). Approximately 91.22% of the paired-end reads were mapped properly.

Differential expression analysis

A total of 14,201 and 279 DEGs were identified corresponding to RF and RB assembly, respectively, in four traits. The DEGs identified through RF assembly were more as compared to RB assembly. The number of upregulated genes was 7,190 and 126 while the downregulated genes were 7,011 and 153, respectively, expressed in RF and RB assembly (Table 1).

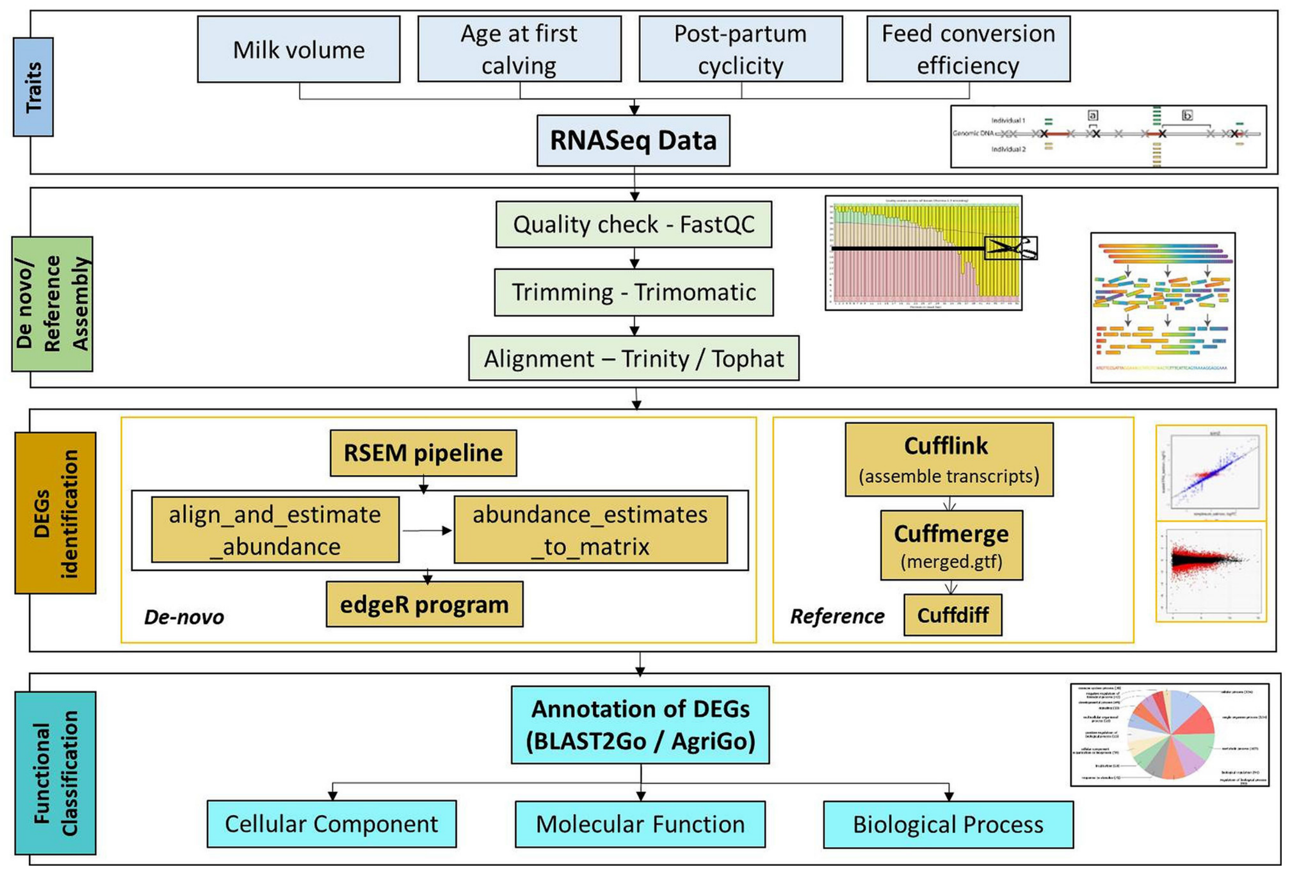


FIGURE 1
Workflow of the RNAseq data analysis.

TABLE 1 Number of differentially expressed genes in reference-free (RF) assembly and reference-based (RB) assembly.

Traits/genes	Reference-free (RF) assembly		Reference-based (RB) assembly	
	Up-regulated	Down-regulated	Up-regulated	Down-regulated
Milk volume (trait 1)	2,020	2,206	29	25
Age at first calving (trait 2)	726	525	37	62
Post-partum cyclicity (trait 3)	2,970	2,498	15	22
Feed conversion efficiency (trait 4)	1,474	1,782	45	44

While making the comparison among the RF upregulated genes across all four traits, only two common DEGs were common between "milk volume" and "feed conversion efficiency," and two DEGs were common between "post-partum cyclicity" and "feed conversion efficiency". A single gene was common between "age at first calving" and "post-partum cyclicity" (Figure 2A). One DEG was found to be common among "milk volume," "post-partum cyclicity," and "feed conversion efficiency" traits (Figure 2B).

No common genes were identified among the four traits. A higher number of common DEGs were found in upregulated and downregulated categories through RB assembly. In total, 3.8% of all DEGs were upregulated in "milk volume," "age at first calving," and "feed conversion efficiency" traits, which were maximum. For traits, viz., "milk volume," "age at first calving," and "post-partum cyclicity," only two common upregulated DEGs were identified in the RB approach (Figure 2C). There were eight downregulated DEGs (7.3%) common in "age at first calving" and "post-partum cyclicity" (Figure 2D), fairly indicating the level of epigenetic regulation with respect to different traits either through DNA methylation and low expression of mRNA or demethylation.

Gene annotation

Gene ontology (GO) terms of identified genes (RF) were obtained using the BLAST2Go v 2.5 tool. The study revealed that one or more GO terms, viz., 30,290, 4,228, 43,142, and 17,097, were assigned to genes for milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency traits, respectively. GO

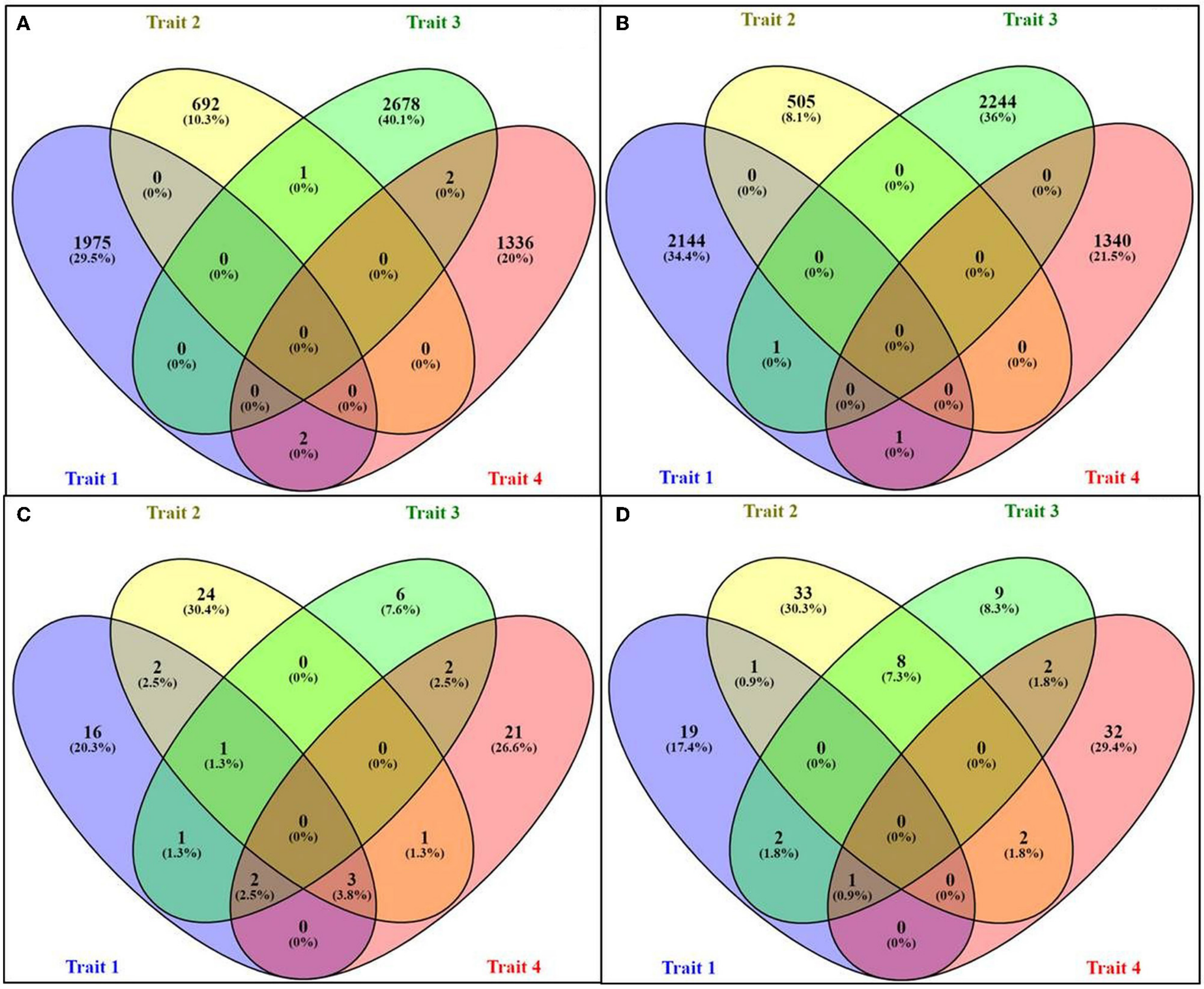


FIGURE 2
Venn diagram representing the number of upregulated and downregulated genes for different traits: **(A)** upregulated genes in RF assembly, **(B)** downregulated genes in RF assembly, **(C)** upregulated genes in RB assembly, and **(D)** downregulated genes in RB assembly.

enrichment analysis classifies gene ontology terms into three broad categories, namely, cellular component, molecular function, and biological process (Figure 3). The binding function (GO: 0005488) was the most represented GO term in the molecular function category followed by cell and its part (GO: 0005623, GO: 0044464) and organelle and its part (GO: 0043226, GO:0044422) as cellular component terms for all the traits. Prominent GO terms that emerged from the RF assembly were similar to the classified terms that emerged from the RB assembly as 2,620 for milk volume, 440 for age at first calving, 3,644 for post-partum cyclicity, and 1,545 for feed conversion efficiency traits (Figure 4).

Variant distribution

Trait-wise distribution of three identified elements as DEGs, SSRs, and SNPs was mapped on the *Bos taurus* genome (Figure 5) and *B. bubalis* genome (Figure 6) using the CIRCOS tool. This depicts the comparative view of the chromosomal-wise distribution of identified elements in various traits, viz., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency. In the RB approach, there are 10,114 SSRs across all four traits, viz., milk volume (3,185), age at first calving (529), post-partum cyclicity (3,828), and feed conversion efficiency (2,572). Mononucleotide SSRs are 6,061 followed by 1,779 for dinucleotide SSRs, 1,417 for trinucleotide SSRs, 55 for tetranucleotide SSRs, 25 for pentanucleotide SSRs, and only two hexanucleotide SSRs. These identified SSRs were further filtered based on their chromosomal locations within the identified DEGs and SNPs. Figure 6 shows the mapping of these resulted in 161 SSRs.

Discussion

The selection programs of domestic animals will be strengthened by a detailed grasp of the genetic variation underlying

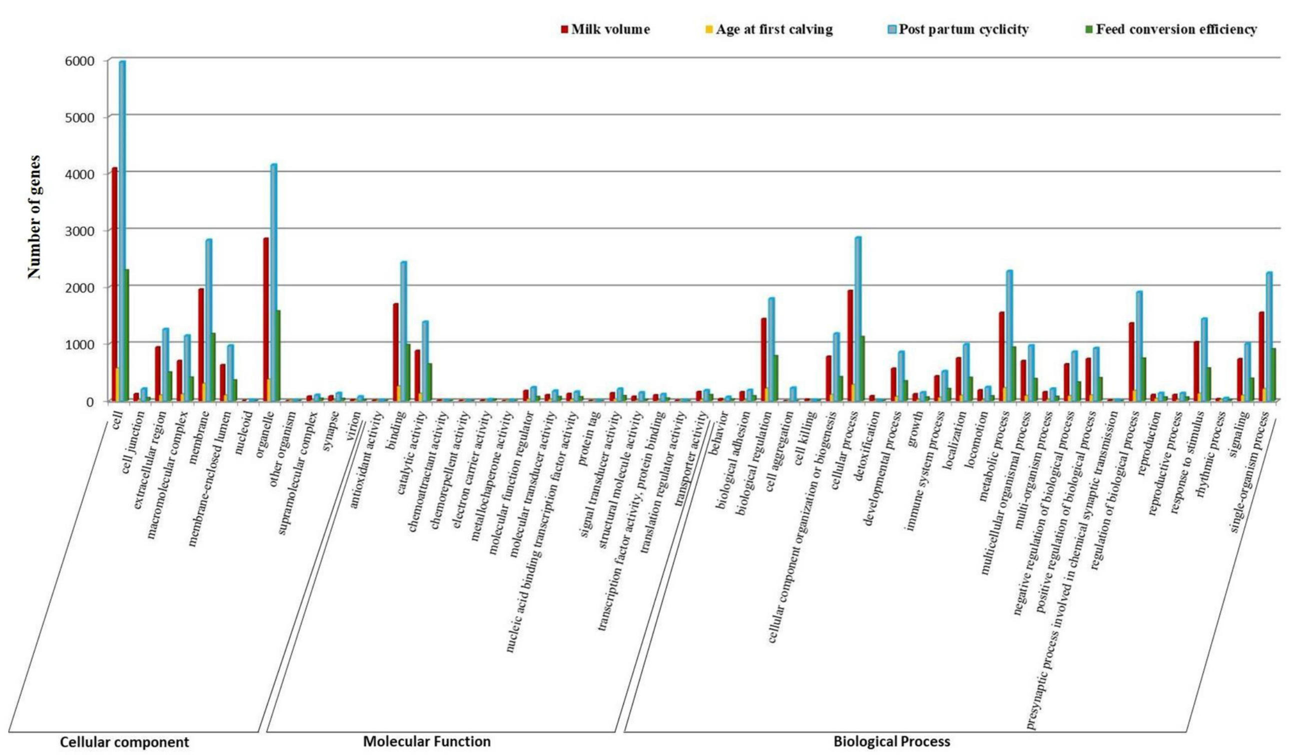


FIGURE 3
Classification of gene ontology terms in reference-free assembly.

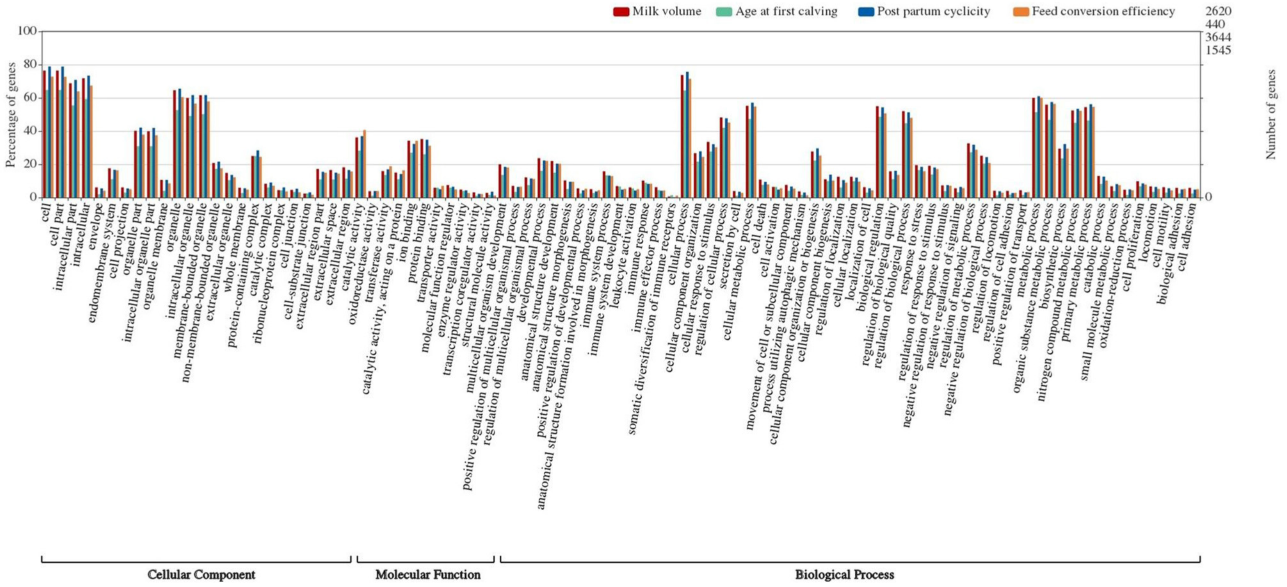


FIGURE 4
Classification of gene ontology terms in reference-based assembly.

complex phenotypes (36) as analyzed in this study. The current study compared the DEGs identified through RF and RB assembly approaches and shows their association with the buffalo traits considered in this study. An attempt was also made to show the chromosomal distribution of DEGs, SNPs, and SSRs in respect of all four considered traits, viz., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency of buffalo using the CIRCOS tool (Figures 5, 6). These maps depict that the maximum number of DEGs are found in post-partum cyclicity, followed by milk volume and feed conversion efficiency in the RF

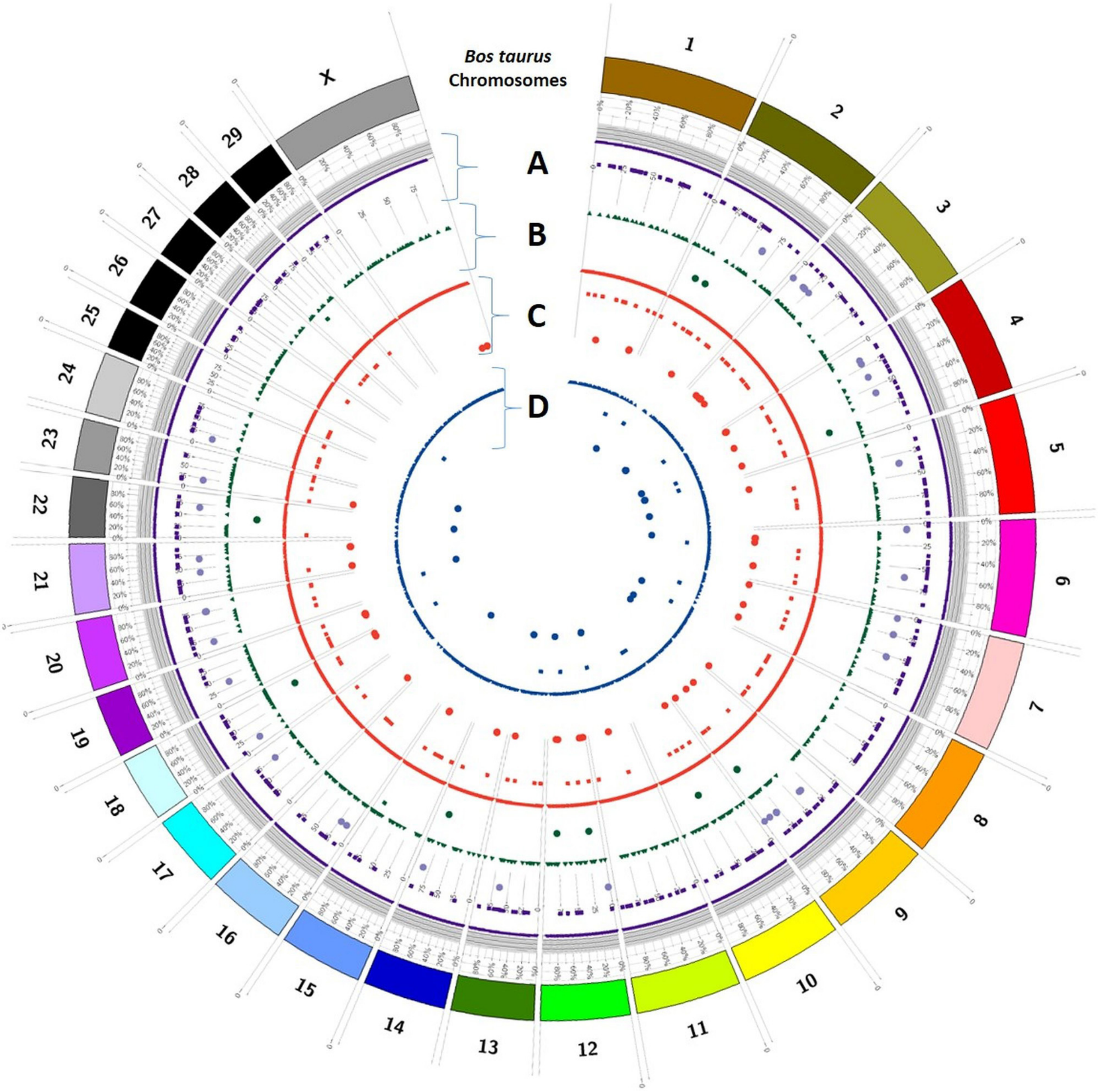


FIGURE 5
Trait-wise distribution of DEGs, corresponding SNPs, and SSRs mapped on the *Bos taurus* chromosomes using CIRCOS [**(A)** milk volume; **(B)** age at first calving; **(C)** post-partum cyclicity; **(D)** feed conversion efficiency].

approach, and age at first calving has the maximum number of DEGs, followed by feed conversion efficiency and milk volume in the RB approach (Table 1).

Phosphatidylinositol 3,4,5 trisphosphate five phosphatase was identified as a common gene that was deregulated for functional diversity in two traits i.e., age at first calving and post-partum cyclicity, indicating the importance of the cell cycle progression in these traits and perhaps regulating the development of embryos (37, 38) (Supplementary Table S1). The FKBP4 gene is responsible for reproductive traits (39, 40). The presence of calcium channel, voltage-dependent, alpha-2/delta subunit 1 (CACNA2D1) gene variant in respect of post-partum cyclicity and feed conversion efficiency traits in our study is crucial for its role in excitation–contraction coupling in neurons, glial cells, and muscle cells (41) (Supplementary Table S1).

The key genes, inositol 1,4,5-triphosphate receptor (ITPR), branched-chain amino acid transaminase (BCAT), and other immunity-related genes, such as T-cell surface glycoprotein and AP-1 transcription factor, are identified as differential genes in our study that are mainly associated with all traits (42). Our study also shows that the identified candidate genes such as growth and hormone receptors, ribosomal proteins, sterol regulatory protein,

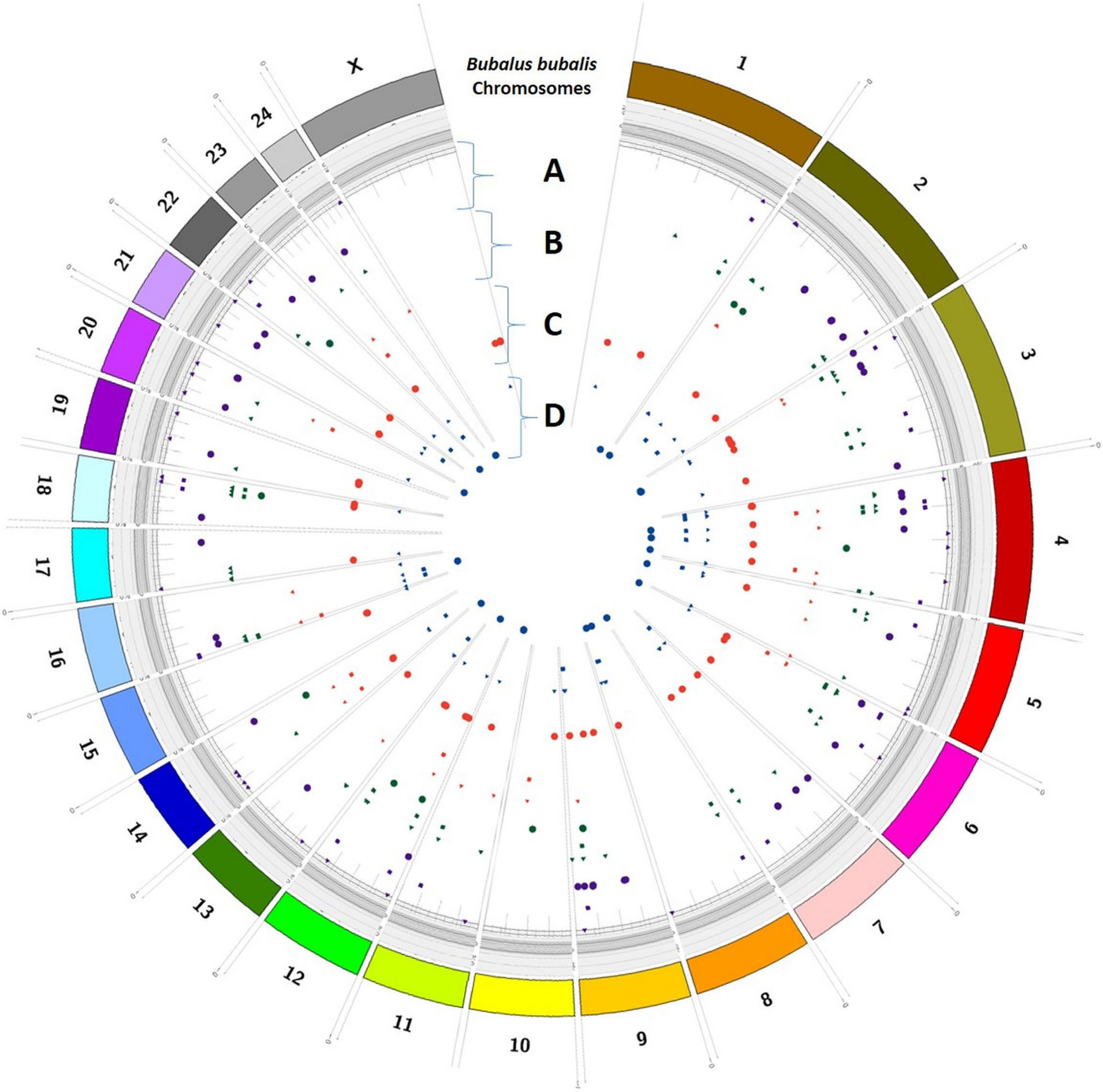


FIGURE 6
Trait-wise distribution of DEGs, corresponding SNPs, and SSRs mapped on the *Bubalus bubalis* chromosomes using CIRCOS [**(A)** milk volume; **(B)** age at first calving; **(C)** post-partum cyclicity; **(D)** feed conversion efficiency].

and GTPase are associated with milk volume as reported earlier by Surya et al. (43), Crisa et al. (44), Wickramasinghe et al. (45), Ma and Corl (46), and Lemay et al. (47), respectively, because of their involvement in histone modification, epidermal differentiation, cell adhesion, and cytoskeletal architecture (48). RNA binding FOX, transmembrane proteins, RNA binding proteins, cytosolic peptidase, and cell adhesion receptor genes were pertinent to age at first calving. These genes are primarily involved in cell proliferation, differentiation, adhesion, the mitotic cycle, DNA replication, RNA transcription, and apoptosis (49) (Supplementary Table S1).

As compared with RF, the RB assembly showed more common genes among the traits. There were a total of 28 genes that were common between different traits in RB compared to RF with only seven common genes (Figure 2). The genes, namely, sphingosine-1-phosphate (SIP), somatostatin (50, 51), BOLA class 1 histocompatibility antigen (52–54), interferon-stimulated/induced protein, and SRY-box transcription factor (55) have maximum homology when aligned with *B. bubalis* genome (UOA_WB_1, Accession GCA_003121395.1; Supplementary Table S2). The gene S1P is a bioactive lipid that acts through cell-surface receptors to promote cell signaling and causes a variety of cellular responses to help in developing immunity against diseases (50, 51). The major histocompatibility complex, such as BOLA class 1 histocompatibility antigen, present in all mammalian species, is crucial for the immune system's development (54). This BOLA histocompatibility complex shows resistance to infectious diseases along with governing the milk volume (52, 53) (Supplementary Table S2). The genes governing disease resistance or susceptibility will positively affect milk productivity.

An immediate defense against viral infection identified in our study is provided by interferon-stimulated genes (ISGs) whose expression is induced by interferon signaling (56). These ISGs also act as potential biomarkers to avoid the occurrence of certain diseases in mammals and eliminate the incidence of adverse reactions to avoid the risk of further damage to the animals (57). This will help in increasing overall productivity that may be either due to an increase in milk volume or due to the application of feed conversion efficiency (Supplementary Table S2).

Post-partum cyclicity-related genes are myosin-related proteins, ribosylhydrolases, and cell adhesion receptors (58) that play a significant role in the growth and immunity of the bovine family. Furthermore, this study also identifies some vital genes (ligand-dependent nuclear receptors such as carboxylase and DLG2) related to feed conversion efficiency for regulating energy homeostasis, apoptosis, immune response, and cell growth in young heifers (59–61) (Supplementary Table S2).

In this study, the enriched GO terms revealed were related to milk production, reproduction, immunological response, and susceptibility/resistance to diseases. GO terms related to milk volume are associated with the biosynthesis of glycoproteins, fatty acids, glycerolipids, sterols, and other biological processes, such as oxidative stress, metabolic processes, transporter activity, divalent metal ion transport, calcium channel activity, acetyltransferase activity, and mRNA processing (Figures 3, 4). This confirms the importance of these processes in lactogenesis (62). The GO terms related to cellular component organization (GO:0071840), cell enzyme activity, or gene expression in response to stimulus (GO:0050896), regulation of biological functions (GO:0050789, GO:0048518, and GO:0048519), single-organism process (GO:0044699), and cell death (GO:0001906) govern physiological processes in animals. These terms are related to age at first calving and feed conversion efficiency (22). Important GO terms stimuli (GO:0050896), regulation of biological quality (GO:0065008), and molecular function (GO:0065009) are related to milk volume and feed conversion efficiency. Genes categorized under cell and cell-part localization (GO:0005623, GO:0030054, and GO:0044464) are prominent in all the considered four traits and are found to regulate different biological processes, viz., trans-membrane transport, regulation of signal transduction, milk production, and chemical transmission (49, 63) (Figures 3, 4). The gene condensing complex subunit 2 (Q3MHQ) encoded by GO term GO:0065007 is linked with the feed efficiency trait. The gene Q3MHQ regulates cell division and improves growth development in an animal by converting interphase chromatin into mitotic chromosomal condensation and is interestingly linked to metabolic pathways involved in feed conversion efficiency (64, 65).

Conclusion

In this study, an attempt has been made to compare reference-free and reference-based approaches to identify and annotate differentially expressed genes in *B. bubalis* for four important traits, viz., milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency. Reference-free (RF) *de novo* transcriptome assembly approach is commonly used due to the non-availability of a complete reference genome with the high-quality genetic information of particular species. In this study, the RF approach identified 7,190 upregulated genes and 7,011 downregulated genes, whereas the RB approach identified 126 and 153 genes, respectively. The number of gene ontology terms associated with identified DEGs for the traits under consideration—milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency—were 30,290, 4,228, 43,142, and 17,097 for the RF approach, compared to 2,620, 440, 3,644, and 1,545 terms for the RB approach. The identified genes and GO terms will establish a sound base for biological postulates which will further improve future animal breeding programs to enhance animal productivity.

Data availability statement

The original transcriptome data presented in the study are publicly available. This data can be found here: NCBI Bioproject, accession number PRJNA934134.

Ethics statement

The animal study was reviewed and approved by Institute Animal Ethics Committee (IAEC) at ICAR-Central Institute for Research on Buffaloes, Hisar, India.

Author contributions

PS, AJ, AB, and IS conducted the experiments. JB and SY did data analysis. DM, JB, KC, and HA conceptualized the study and manuscript. All authors reviewed and approved the final manuscript.

Funding

This research was funded by the project Network Project for Agricultural Bioinformatics and Computational Biology under CABin Scheme, ICAR (F.no. Agril. Edn.4-1/2013-A&P).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2023.1160486/full#supplementary-material

References

1. Warriach H, McGill D, Bush R, Wynn P, Chohan K. A review of recent developments in buffalo reproduction—a review. *Asian-Australas J Anim Sci.* (2015) 28:451. doi: 10.5713/ajas.14.0259

2. Niranjan S, Goyal S, Dubey P, Kumari N, Mishra S, Mukesh et al. Genetic diversity analysis of buffalo fatty acid synthase (FASN) gene and its differential expression among bovines. *Gene.* (2016) 575:506–12 doi: 10.1016/j.gene.2015.09.020

3. Jerome A, Bhati J, Mishra DC, Chaturvedi KK, Rao AR, Rai A, et al. MicroRNA-related markers associated with corpus luteum tropism in buffalo (*Bubalus bubalis*). *Genomics.* (2020) 112:108–13. doi: 10.1016/j.ygeno.2019.01.018

4. Mishra DC, Sikka P, Yadav S, Bhati J, Paul SS, Jerome A, et al. Identification and characterization of trait-specific SNPs using ddRAD sequencing in water buffalo. *Genomics.* (2020) 112:3571–8. doi: 10.1016/j.ygeno.2020.04.012

5. El-Salam A, Mohamed H, El-Shibiny S. A comprehensive review on the composition and properties of buffalo milk. *Dairy Sci Tech.* (2011) 91:663–99. doi: 10.1007/s13594-011-0029-2

6. Guzman JLG, Lázaro SF, do Nascimento AV, de Abreu Santos DJ, Cardoso DF, Becker Scalez DC, et al. Genome-wide association study applied to type traits related to milk yield in water buffaloes (*Bubalus bubalis*). *J Dairy Sci.* (2020) 103:1642–50. doi: 10.3168/jds.2019-16499

7. Eastham NT, Coates A, Cripps P, Richardson H, Smith R, Oikonomou G. Associations between age at first calving and subsequent lactation performance in UK Holstein and Holstein-Friesian dairy cows. *PLoS ONE.* (2018) 13:e0197764. doi: 10.1371/journal.pone.0197764

8. Japheth KP, Kumaresan A, Patbandha TK, Baithalu RK, Selvan AS, Nag P, et al. Supplementation of a combination of herbs improves immunity, uterine cleansing and facilitate early resumption of ovarian cyclicity: a study on post-partum dairy buffaloes. *J Ethnopharm.* (2021) 272:113931. doi: 10.1016/j.jep.2021.113931

9. Sethi M, Shah N, Mohanty TK, Bhakat M, Dewry RK, Yadav DK, et al. The induction of cyclicity in postpartum anestrus buffaloes: a review. *J Exp Zool.* (2021) 24:989–97.

10. Sikka P, Nath A, Paul SS, Andonissamy J, Mishra DC, Rao AR, et al. Inferring relationship of blood metabolic changes and average daily gain with feed conversion efficiency in murrah heifers: machine learning approach. *Front Vet Sci.* (2020) 7:518. doi: 10.3389/fvets.2020.00518

11. Haldar A, Prakash BS. Effects of growth hormone-releasing factor on growth hormone response, growth and feed conversion efficiency in buffalo heifers (*Bubalus bubalis*). *Vet J.* (2007) 174:384–9. doi: 10.1016/j.tvjl.2006.10.003

12. Reddy RM. Feed-milk conversion efficiency of dairy animals and methane emissions across different landholding groups. *International J Sci Res.* (2020) 10:1242–8. doi: 10.21275/SR21527174803

13. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat Commun.* (2021) 12:5647. doi: 10.1038/s41467-021-25805-y

14. Mishra DC, Smita S, Singh I, Devi MN, Kumar S, Farooqi MS, et al. Prediction of novel putative miRNAs and their targets in buffalo. *Indian J Anim Sci.* (2017) 87:59–63. doi: 10.56093/ijans.v87i1.66861

15. Pokharel K, Peippo J, Honkatukia M, Seppälä A, Rautiainen J, Ghanem N, et al. Integrated ovarian mRNA and miRNA transcriptome profiling characterizes the genetic basis of prolificacy traits in sheep (*Ovis aries*). *BMC Genomics.* (2018) 19:104. doi: 10.1186/s12864-017-4400-4

16. Liu Y, Wu X, Xie J, Wang W, Xin J, Kong F, et al. Identification of transcriptome differences in goat ovaries at the follicular phase and the luteal phase using an RNA-Seq method. *Theriogenology.* (2020) 158:239–49. doi: 10.1016/j.theriogenology.2020.06.045

17. Gong X, Zheng M, Zhang J, Ye Y, Duan M, Chamba Y, et al. Transcriptomics-based study of differentially expressed genes related to fat deposition in Tibetan and yorkshire pigs. *Front Vet Sci.* (2022) 9:919904. doi: 10.3389/fvets.2022.919904

18. El Nahas SM, Mossallam AAA. AcuI identifies water buffalo CSN3 genotypes by RFLP analysis. *J Genet.* (2014) 93:e94–6. doi: 10.1007/s12041-014-0427-3

19. Paraguison RC, Faylon MP, Flores EB, Cruz LC. Improved RAPD-PCR for discriminating breeds of water buffalo. *Biochem Genet.* (2012) 50:579–84. doi: 10.1007/s10528-012-9502-8

20. Mishra DC, Yadav S, Sikka P, Jerome A, Paul SS, Rao AR, et al. SNPRBb: economically important trait specific SNP resources of buffalo (*Bubalus bubalis*). *Conserv Genet Resour.* (2021) 13:283–9. doi: 10.1007/s12686-021-01210-x

21. Gargani M, Pariset L, Soysal MI, Ozkan E, Valentini A. Genetic variation and relationships among Turkish water buffalo populations. *Animal Genet.* (2010) 41:93–6. doi: 10.1111/j.1365-2052.2009.01954.x

22. Pathak RK, Kim JM. Vetinformatics from functional genomics to drug discovery: insights into decoding complex molecular mechanisms of livestock systems in veterinary science. *Front Vet Sci.* (2022) 9:1008728. doi: 10.3389/fvets.2022.1008728

23. Deng T, Liang A, Liang S, Ma X, Lu X, Duan A, et al. Integrative analysis of transcriptome and GWAS data to identify the hub genes associated with milk yield trait in buffalo. *Front Genet.* (2019) 10:36. doi: 10.3389/fgene.2019.00036

24. Liu S, Ye T, Li Z, Li J, Jamil AM, Zhou Y, et al. Identifying hub genes for heat tolerance in water buffalo (*Bubalus bubalis*) using transcriptome data. *Front Genet.* (2019) 10:209. doi: 10.3389/fgene.2019.00209

25. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE.* (2012) 7:e30619. doi: 10.1371/journal.pone.0030619

26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

27. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data.* (2010). Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed March 30, 2023).

28. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech.* (2011) 29:644–52. doi: 10.1038/nbt.1883

29. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* (2013) 8:1494–512. doi: 10.1038/nprot.2013.084

30. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* (2013) 14:R36. doi: 10.1186/gb-2013-14-4-r36

31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* (2012) 9:357–9. doi: 10.1038/nmeth.1923

32. R Core Team. (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Available online at: http://www.R-project.org/ (accessed March 30, 2023).

33. Robinson MD, Mccarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616

34. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* (2011) 12:323. doi: 10.1186/1471-2105-12-323

35. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* (2008) 36:3420–35. doi: 10.1093/nar/gkn176

36. Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc Royal Soc B.* (2016) 283:20160569. doi: 10.1098/rspb.2016.0569

37. Qiao F, Law HC, Krieger KL, Clement EJ, Xiao Y, Buckley SM, et al. Ctdp1 deficiency leads to early embryonic lethality in mice and defects in cell cycle progression in MEFs. *Biol Open.* (2021) 10:bio057232. doi: 10.1242/bio.057232

38. Dubon MAC, Pedrosa VB, Feitosa FLB, Costa RB, de Camargo GMF, Silva MR, et al. Identification of novel candidate genes for age at first calving in nellore cows using

a SNP chip specifically developed for *Bos taurus* indicus cattle. *Theriogenology.* (2021) 173:156–62. doi: 10.1016/j.theriogenology.2021.08.011

39. Hulsegge I, Woelders H, Smits M, Schokker D, Jiang L, Sorensen P. Prioritization of candidate genes for cattle reproductive traits, based on protein-protien interactions, gene expression, and text-mining. *Physiol Genomics.* (2013) 45:400–6. doi: 10.1152/physiolgenomics.00172.2012

40. Wathes DC, Cheng Z, Chowdhury W, Fenwick MA, Fitzpatrick R, Morris DG, et al. Negative energy balance alters global gene expression and immune responses in the uterus of postpartum dairy cows. *Physiol Genomics.* (2009) 39:1–13. doi: 10.1152/physiolgenomics.00064.2009

41. Magotra A, Gupta ID, Verma A, Alex R, Arya A, Vineeth MR, et al. Characterization and validation of point mutation in Exon 19 of Calcium channel, voltage-dependent, Alpha-2/Delta subunit 1(CACNA2D1) gene and its relationship with mastitis traits in Sahiwal. *Indian J Anim Res.* (2018) 52:61–4. doi: 10.18805/ijar.10990

42. Zheng X, Ju Z, Wang J, Li Q, Huang J, Zhang A, et al. Single nucleotide polymorphisms, haplotypes and combined genotypes of *LAP3* gene in bovine and their association with milk production traits. *Mol Biol Rep.* (2011) 38:4053–61. doi: 10.1007/s11033-010-0524-1

43. Surya T, Vineeth MR, Sivalingam J, Tantia MS, Dixit SP, Niranjan SK, et al. Genomewide identification and annotation of SNPs in *Bubalus bubalis*. *Genomics.* (2019) 111:1695–8. doi: 10.1016/j.ygeno.2018.11.021

44. Crisa A, Ferre F, Chillemi G, Moioli B. RNA-Sequencing for profiling goat milk transcriptome in colostrum and mature milk. *BMC Vet Res.* (2016) 12:264. doi: 10.1186/s12917-016-0881-7

45. Wickramasinghe S, Rincon G, Islas-Trejo A, Medrano JF. Transcriptional profiling of bovine milk using RNA sequencing. *BMC Genomics.* (2012) 13:45. doi: 10.1186/1471-2164-13-45

46. Ma L, Corl BA. Transcriptional regulation of lipid synthesis in bovine mammary epithelial cells by sterol regulatory element binding protein-1. *J Dairy Sci.* (2012) 95:3743–55. doi: 10.3168/jds.2011-5083

47. Lemay DG, Lynn DJ, Martin WF, Neville MC, Casey TM, Rincon G, et al. The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biol.* (2009) 10:R43. doi: 10.1186/gb-2009-10-4-r43

48. Du C, Deng T, Zhou Y, Ye T, Zhou Z, Zhang S, et al. Systematic analyses for candidate genes of milk production traits in water buffalo (*Bubalus bubalis*). *Anim Genet.* (2019) 50:207–16. doi: 10.1111/age.12739

49. Fan H, Wu Y, Qi X, Zhang J, Li J, Gao X, et al. Genome-wide detection of selective signatures in Simmental cattle. *J Appl Genet.* (2014) 55:343–51. doi: 10.1007/s13353-014-0200-6

50. Pyne NJ, Pyne S. Sphingosine 1-phosphate receptor 1 signaling in mammalian cells. *Molecules.* (2017) 22:344. doi: 10.3390/molecules22030344

51. Spiegel S, Milstien S. Sphingosine-1-phosphate: an enigmatic signalling lipid. *Nat Rev Mol Cell Biol.* (2003) 4:397–407. doi: 10.1038/nrm1103

52. Rupp R, Hernandez A, Mallard BA. Association of bovine leukocyte antigen (BoLA) DRB3.2 with immune response, mastitis, and production and type traits in *Canadian holsteins. J Dairy Sci.* (2007) 90:1029–38. doi: 10.3168/jds.S0022-0302(07)71589-8

53. Pashmi M, Qanbari S, Ghorashi SA, Sharifi AR, Simianer H. Analysis of relationship between bovine lymphocyte antigen DRB3.2 alleles, somatic cell count and milk traits in Iranian holstein p. *J Anim Breed Genet.* (2009) 126:296–303. doi: 10.1111/j.1439-0388.2008.00783.x

54. Behl JD, Verma NK, Tyagi N, Mishra P, Behl R, Joshi BK. The major histocompatibility complex in bovines: a review. *ISRN Vet Sci.* (2012) 28:872710. doi: 10.5402/2012/872710

55. Lefebvre V, Dumitriu B, Penzo-Méndez A, Han Y, Pallavi B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Int J Biochem Cell Biol.* (2007) 39:2195–214. doi: 10.1016/j.biocel.2007.05.019

56. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol.* (2017) 15:e2004086. doi: 10.1371/journal.pbio.2004086

57. Zheng Z, Wang L, Pan J. Interferon-stimulated gene 20-kDa protein (ISG20) in infection and disease: review and outlook. *Intractable Rare Dis Res.* (2017) 6:35–40. doi: 10.5582/irdr.2017.01004

58. Abo-Ismail MK, Brito LF, Miller SP, Sargolzaei M, Grossi DA, Moore SS, et al. Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet Sel Evol.* (2017) 49:82. doi: 10.1186/s12711-017-0356-8

59. Sasago N, Abe T, Sakuma H, Kojima T, Uemoto Y. Genome-wide association study for carcass traits, fatty acid composition, chemical composition, sugar, and the effects of related candidate genes in Japanese black cattle. *Anim Sci J.* (2017) 88:33–44. doi: 10.1111/asj.12595

60. Sun C, Southard C, Witonsky DB, Kittler R, Di Rienzo A. Allele-specific down-regulation of *RPTOR* expression induced by retinoids contributes to climate adaptations. *PLoS Genet.* (2010) 6:e1001178. doi: 10.1371/journal.pgen.1001178

61. Setoguchi K, Watanabe T, Weikard R, Albrecht E, Kuhn C, Kinishita A, et al. The SNPc.1326T>G in the non-SMC condensin I complex, subunit G(NCAPG) gene encoding a p.Ile442Met variant is associated with an increase in body frame size at puberty in cattle. *Animal Genet.* (2011) 42:650–5. doi: 10.1111/j.1365-2052.2011.02196.x

62. Lemay DG, Neville MC, Rudolph MC, Pollard KS, German JB. Gene regulatory networks in lactation: identification of global principles using bioinformatics. *BMC Syst Biol.* (2007) 1:56. doi: 10.1186/1752-0509-1-56

63. Nayeri S, Stothard P. Tissues, metabolic pathways and genes of key importance in lactating dairy cattle. *Springer Sci Rev.* (2016) 4:49–77. doi: 10.1007/s40362-016-0040-3

64. Daetwyler H, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet.* (2014) 46:858–65. doi: 10.1038/ng.3034

65. Widmann P, Reverter A, Weikard R, Suhre K, Hammon HM, Albrecht E, et al. Systems biology analysis merging phenotype, metabolomic and genomic data identifies non-SMC condensin I complex, subunit G (NCAPG) and cellular maintenance processes as major contributors to genetic variability in bovine feed efficiency. *PLoS ONE.* (2015) 10:e0124574. doi: 10.1371/journal.pone.0124574

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 21 June 2023
DOI 10.3389/fvets.2023.1174700

Check for updates

OPEN ACCESS

EDITED BY
Sabine G. Gebhardt-Henrich,
University of Bern, Switzerland

REVIEWED BY
Eamonn John Keogh,
University of California, Riverside,
United States
Alireza Abdoli,
University of California,
Riverside, United States

*CORRESPONDENCE
Pin Jern Ker
✉ pinjern@uniten.edu.my

RECEIVED 27 February 2023
ACCEPTED 19 May 2023
PUBLISHED 21 June 2023

CITATION
Bakar MAAA, Ker PJ, Tang SGH, Baharuddin MZ, Lee HJ and Omar AR (2023) Translating conventional wisdom on chicken comb color into automated monitoring of disease-infected chicken using chromaticity-based machine learning models.
Front. Vet. Sci. 10:1174700.
doi: 10.3389/fvets.2023.1174700

Translating conventional wisdom on chicken comb color into automated monitoring of disease-infected chicken using chromaticity-based machine learning models

Mohd Anif A. A. Bakar[1], Pin Jern Ker[1]*, Shirley G. H. Tang[2], Mohd Zafri Baharuddin[1], Hui Jing Lee[3] and Abdul Rahman Omar[4,5]

[1]Department of Electrical and Electronics Engineering, College of Engineering, Institute of Sustainable Energy, Universiti Tenaga Nasional, Kajang, Malaysia, [2]Center for Toxicology and Health Risk Studies (CORE), Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia, [3]Department of Electrical and Electronics Engineering, College of Engineering, Institute of Power Engineering, Universiti Tenaga Nasional, Kajang, Malaysia, [4]Department of Veterinary Pathology and Microbiology, Faculty of Veterinary, Universiti Putra Malaysia, Serdang, Malaysia, [5]Institute of Bioscience, Universiti Putra Malaysia, Serdang, Malaysia

Bacteria- or virus-infected chicken is conventionally detected by manual observation and confirmed by a laboratory test, which may lead to late detection, significant economic loss, and threaten human health. This paper reports on the development of an innovative technique to detect bacteria- or virus-infected chickens based on the optical chromaticity of the chicken comb. The chromaticity of the infected and healthy chicken comb was extracted and analyzed with International Commission on Illumination (CIE) XYZ color space. Logistic Regression, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and Decision Trees have been developed to detect infected chickens using the chromaticity data. Based on the X and Z chromaticity data from the chromaticity analysis, the color of the infected chicken's comb converged from red to green and yellow to blue. The development of the algorithms shows that Logistic Regression, SVM with Linear and Polynomial kernels performed the best with 95% accuracy, followed by SVM-RBF kernel, and KNN with 93% accuracy, Decision Tree with 90% accuracy, and lastly, SVM-Sigmoidal kernel with 83% accuracy. The iteration of the probability threshold parameter for Logistic Regression models has shown that the model can detect all infected chickens with 100% sensitivity and 95% accuracy at the probability threshold of 0.54. These works have shown that, despite using only the optical chromaticity of the chicken comb as the input data, the developed models (95% accuracy) have performed exceptionally well, compared to other reported results (99.469% accuracy) which utilize more sophisticated input data such as morphological and mobility features. This work has demonstrated a new feature for bacteria- or virus-infected chicken detection and contributes to the development of modern technology in agriculture applications.

KEYWORDS

machine learning, classification model, chromaticity, agriculture, chicken comb, image processing, diseases-infected chicken, energy

1. Introduction

The increase in human population has forced poultry meat production to increase (1). However, mass production in the poultry industry may be more vulnerable to disease outbreaks in farmed animals due to the increased number of animals per area and prolonged usage of antibiotics (2). The World Bank reported a direct cost of $20 billion for disease outbreak events between 1988 to 2006 (3), including public and animal health costs, compensation, production, and revenue costs. Plus, indirect losses, including animal product chain, trade, and tourism, were estimated to be more than $200 billion worldwide (3). For instance, the United States and China's poultry industry recorded huge economic losses and threats to human health due to several poultry-related diseases such as the H7N9 avian influenza virus outbreak in 2013 (4), multistate foodborne outbreak of *Salmonella* Typhimurium (5), avian influenza outbreaks in 2022 (6, 7), foodborne pathogens such as *Campylobacter*, *Escherichia coli*, *Salmonella*, and Norovirus (8), severe respiratory illness among poultry slaughter plant workers due to *Chlamydia psittaci* (9), and human infection with the influenza A (H5N6) virus of avian origin (10). Although the viruses are preventable, curable, and controllable, there is still a continuous threat that they could start a pandemic if the viruses develop the ability to spread among humans effectively. Therefore, early detection of diseases in poultry production is a primary concern to prevent a major outbreak that would affect the economy and human health.

Numerous disease detection methods have been proposed, developed, and widely applied to give early detection to prevent this catastrophe. The conventional method of detecting infected chicken was using physical examination and laboratory tests. The physical examination is a way of seeing infected chicken through observation of clinical signs or changes in behavior and physical appearance of the chicken individually. The suspected chicken will be evicted from the flocks and undergo laboratory tests such as culture (11–13), polymerase chain reaction (PCR) (14, 15), enzyme-linked immunosorbent assay (ELISA) (16, 17) and lateral flow assay (LFA) (18, 19). Biological samples such as blood, cloacal swabs, organs, and feces were collected from the suspected chicken for the test. Apart from the requirement of trained personnel to conduct the tests, these methods are considered costly due to the equipment needed, such as a thermocycler, ELISA reader, PCR buffer, syringe, swab kit, and petri dish for sampling and detecting the pathogen (20). Overall, these methods can detect infected chickens with high precision and specificity. However, many other factors, such as cost and time taken for detection, were compromised, which makes it almost impossible to be implemented, especially for large-scale poultry producers.

The rapid development of modern technology has introduced the development of biosensors to detect infections with consideration of other factors such as sensitivity, cost, efficiency, and time taken for detection (21–23). Although biosensors can detect infected chickens faster than laboratory tests with good sensitivity and accuracy, each method was considered intrusive due to the biological sample needed for the test. Non-intrusive and non-invasive techniques in giving an early warning for detecting infected chickens based on their vocalization, video, and image have been introduced with the aid of advanced information technologies, especially machine learning. Several researchers have successfully detected infected chickens based on their abnormal sounds like rales, sneezing, and coughing (24–26). However, it was challenging to detect infected chickens individually based on vocalization because more than one chicken may sneeze or cough simultaneously. Computer vision, like digital images and video, can detect and classify infected chickens in real-time, and many different methods have been proposed (27–31). However, these works carried out the classification based on locomotor and mobility of the chicken (27), differences in morphological features (28), differences in posture and feather images (29), using an abnormal swelling image (30), and the correlation of the optical flow parameters with the occurrence of hockburn in chicken (31).

In conventional understanding, the infected chicken can be detected based on the biological change in the appearance of the chicken itself, especially its comb. For example, the Newcastle disease infection would show clinical signs such as swollen comb (32), nodular lesions on its comb characterized by fowl pox disease infection (33), and fatty liver hemorrhagic syndrome would show clinical signs of a pale comb (34). Previous studies have reported on the relationship between comb color and size with the immunity system of birds using spectrophotometry (35, 36). However, these results were based upon data from red grouse (bird) combs and it is still unclear on the correlation between the comb's chromaticity and bacteria- or virus-infection, since these works were investigating only the immunity system of the birds. To the best of our knowledge, there is no specific research work that correlates the optical chromaticity of the chicken comb with infectious diseases using image processing. Therefore, this work investigates on the effectiveness of utilizing image processing techniques incorporated with machine learning algorithms to correlate the color of the chicken comb with bacteria- or virus-infected chicken. The difference between infected and healthy chicken comb is analyzed based on chromaticity data. Since computer is a low-cost, non-invasive and non-intrusive method for detecting infected chicken, digital image colorimetry was adopted in this work. Using the chromaticity data, machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors, and Decision Tree, were developed to classify the infected and healthy chickens. Each model's performance, advantages, and disadvantages for this current application were analyzed in this study.

2. Image processing and machine learning algorithms

A digital image is a combination of color space data, and many researchers had performed colorimetry studies based on digital image color space data for a few applications and areas (37–41). Since digital image colorimetry is a well-known method for describing perceived color, this technique was used to extract the color component of the chicken comb at pixel level and the average pixel color component bounded on the comb area. The Red Green Blue triplets, RGB values were extracted, normalized, and linearly transformed into CIE XYZ color space using the developed Python program and ImageJ software. Normalized CIE XYZ, named the CIE *xyz* component, was studied and analyzed incorporated with the machine learning model, Logistic Regression. The supervised machine learning classification algorithms, Logistic Regression, SVM with different types of kernels, KNN, and Decision Tree model were used to classify the chicken health based on the color component. The models were trained and validated to analyze the performance parameter in this current application. Figure 1 shows the workflow of this study, from the RGB color data extraction methods to the chromaticity data analysis and the

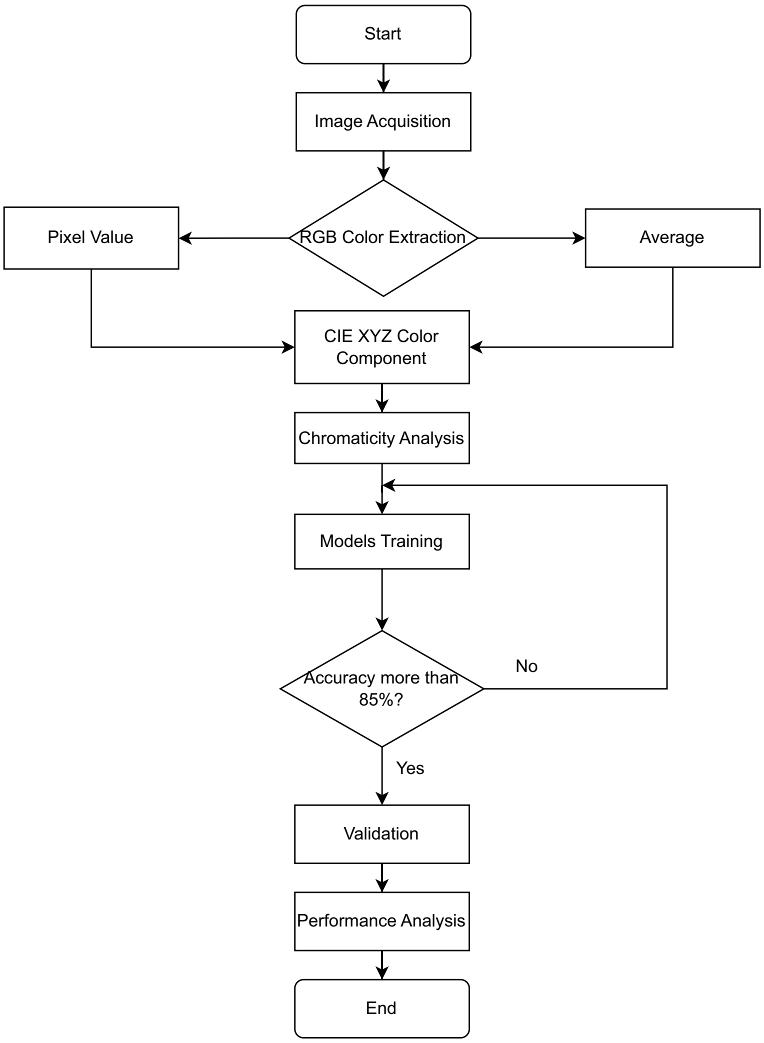


FIGURE 1
Workflow of the RGB color data extraction methods, data analysis and development of machine learning models.

development of machine learning models to classify chicken health. The details for the major stage of the method, which are image acquisition, data organization, image processing and data labeling, CIE XYZ color space, supervised machine learning algorithms, and performance parameter, are discussed in the following subsections.

2.1. Image acquisition

Digital image data were manually collected from various sources such as journals, short communications, articles, veterinary websites, and blogs through the open-source Google Search engine because no specified image dataset related to this work could be obtained. A total of 122 images were downloaded and classified into two groups, healthy and infected chickens, with 61 images in each group without considering any specific quality such as resolution, lighting condition, the pixel value of the image, the distance between the camera and the chicken, and the angle view of the chicken. The images were labeled as healthy and infected based on the source's justification. All the image data including masked chicken comb images and sources have been uploaded to a GitHub repository.[1] All chickens were assumed to be alive based on general observation. Images were selected based on the feather color to indicate a type of chicken, and the current work considered chickens with white feathers only. However, the chicken husbandry care such as the diet, age, temperature, humidity of surroundings, and severity of the diseases were not considered in this work. As presented in Figure 2, most of the chickens in the infected class dataset were infected with Newcastle disease (25%), followed by infectious bronchitis (10%) and avian influenza (8%).

2.2. Data organization

The image data was split into training and validation sets to reduce bias in training the model. Eighty images were randomly picked as a training dataset for fitting the models, and the remaining images were

[1] https://github.com/anifakhmal/Infected-vs-Healthy-chick.git

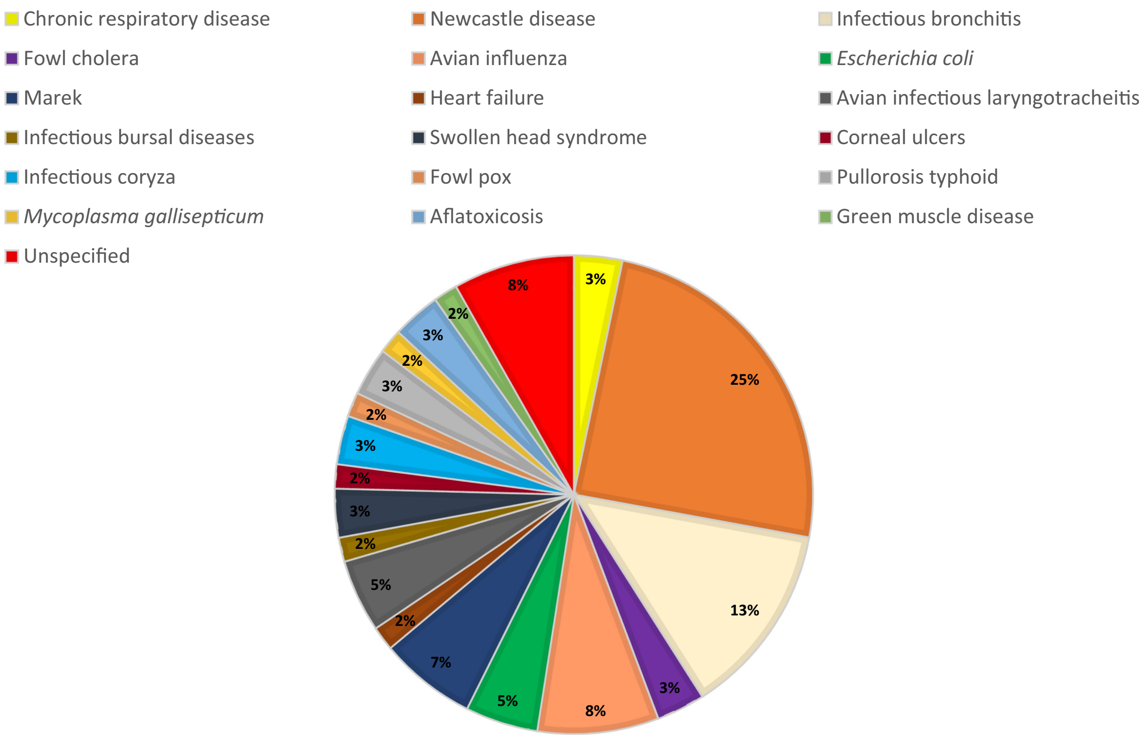


FIGURE 2
Percentage distribution of different diseases for the infected chicken image dataset.

used as the validation set. The models were validated by 42 healthy and infected chickens, which were randomly distributed but properly structured to represent all diseases. For infected chicken with a total of only 2 or 3 images, such as chronic respiratory diseases, fowl cholera, infectious coryza, swollen head syndrome, aflatoxicosis, *E. coli*, avian infectious laryngotracheitis, and pullorosis typhoid, one image was randomly picked from each group for validation. Two photos were selected for validation from each disease group containing 4 to 8 images representing Marek, avian influenza, infectious bronchitis, and unspecified diseases. The most considerable portion of the validation dataset belongs to Newcastle disease, with 23.81% (5 images out of 21 total) due to overall image acquisition. However, infectious bursal disease, *Mycoplasma gallisepticum*, heart failure, fowlpox, corneal ulcers, and green muscle disease were not included in the validation dataset, due to a lack of image data. Overall, the training dataset consists of 40 healthy and 40 infected chicken images, while the validation set consists of 21 healthy and 21 infected chicken images.

2.3. Image processing and data labeling

The raw image data were not uniform in size and resolution. The image of the chicken head was cropped manually to analyze its comb color within the comb area excluding the region that has overlayed text. This work used two methods to extract the RGB value of the chicken comb. The first method was by extracting 3 RGB sample points within the area of the chicken comb, as shown in Figure 3A. The second method was by extracting the average RGB value of all pixels within the chicken comb, as shown in Figure 3B. Throughout this paper, the first method will be named the pixel-level method, and the second method will be named the pixel-averaging method.

Figure 3A shows that three sample points were taken from the image at coordinates (70.36), (127.34), and (167.56). The image coordinate was specified based on the chicken comb using the convention of width and height. Figure 3B shows that the chicken comb was manually selected to calculate the average value of all the extracted RGB values within the selected region. The RGB data was normalized and transformed into CIE XYZ color space which was discussed theoretically in the next subsection. The collected RGB and CIE XYZ color space data were saved in Macintosh (.csv) format for further analysis. The infected chicken was labeled as 0 for the true positive event, and the healthy chicken was labeled as 1 for the true negative event as described in the literature (42).

2.4. CIE XYZ color space

In this work, the CIE XYZ color space (43) was utilized to analyze the chromaticity of the infected and healthy chicken combs. The extracted RGB data were normalized and converted to CIE XYZ color space using the linear matrix transformation as shown in Equation 1. The formula was directly adopted from (43) the Rec. 709 RGB standards with its reference D65 white point for all images.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124564 & 0.3575761 & 0.1804375 \\ 0.2126729 & 0.7151522 & 0.0721750 \\ 0.0193339 & 0.1191920 & 0.9503041 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

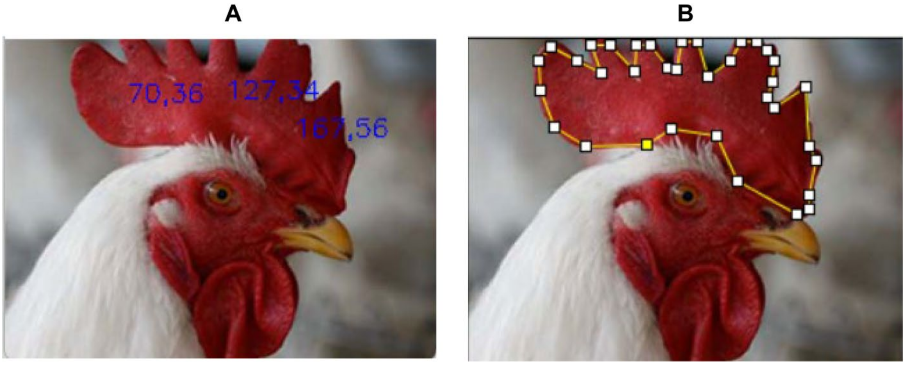


FIGURE 3
Data extraction using **(A)** pixel-level method and **(B)** pixel-averaging method.

The XYZ values were normalized to restrict the range from 0 to 1 and denoted as *x*, *y*, and *z* values. The formulas used in normalizing the value were expressed in Equations (2)–(4).

$$x = \frac{X}{X + Y + Z} \quad (2)$$

$$y = \frac{Y}{X + Y + Z} \quad (3)$$

$$z = \frac{Z}{X + Y + Z} \quad (4)$$

The scatter plots of *xy*, *yz*, and *xz* values were analyzed to determine the differences between the healthy and infected chickens.

2.5. Supervised machine learning algorithms

This research work utilized four different classifier algorithms, namely Logistic Regression, SVM, KNN, and Decision Tree. Scikit Learn library was used for pre-processing the data and training the models as specified in the package (44). The *x* and *y* chromaticity data were utilized as the features for the classifier. The chromaticity data features were standardized using the StandardScaler module from the Scikit library for faster convergence and better results. Further fundamental, theoretical, and mathematical theories of these models were well discussed in the library documentation. Hyperparameters of each model were adjusted and the best model was selected and discussed based on the confusion matrix, which was discussed in the performance parameter subsection. The advantages and disadvantages of deploying each model were also addressed for this current application in section 3.2.

2.5.1. Logistic regression

The logistic regression model is a supervised machine learning model to predict the class probability, which ranges from 0 to 1 in our application. The model predicts 0 for probability ranging from 0 to 0.5, and the class belongs to the positive event or infected chicken. The theory of the logistic regression model was explained in literature (45). The logistic function was used to restrict the linear regression model's output to a range from 0 to 1. The general logistic equation is given in Equation 5. Note that, $p(y)$ is the function for the probability value, and variable y in the equation corresponds to the input function for the logistic equation.

$$p(y) = \frac{1}{1 + e^{-y}} \quad (5)$$

Since the logistic regression was restricting the linear regression model, the final equation for the model is stated in Equation 6.

$$p\left(f\left(x_1,x_2\right)\right) = \frac{1}{1 + e^{-\left(B_0 + B_1 x_1 + B_2 x_2\right)}} \quad (6)$$

where $f\left(x_1,x_2\right)$ is the sigmoid input function for the logistic equation, x_1 and x_2 correspond to the predictor or chromaticity data for the classifier, and B_0, B_1, and B_2 correspond to the coefficient of the predictors. Current work will utilize the sigmoid input function $f\left(x_1,x_2\right)$, to analyze and correlate the chromaticity data and the health status of the chickens. The general function is stated in Equation 7.

$$f\left(x_1,x_2\right) = B_0 + B_1 x_1 + B_2 x_2 \quad (7)$$

The iteration of the cost function, C parameter, was carried out and the highest accuracy performance was analyzed.

2.5.2. Support vector machine

SVMs are a popular supervised learning technique for outliers' detection, regression, and classification. SVM algorithms take data as input and transform it into the desired form using a set of mathematical functions referred to as the kernel. Given that the ScikitLearn library offers four distinct kernel functions (44)—Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid—the current work will develop the models across all four kernels. The Linear, Polynomial, RBF, and Sigmoid kernel functions are given in Equations (8)–(11), respectively.

$$K\left(x_1,x_2\right) = x_1.x_2 \tag{8}$$

$$K\left(x_1,x_2\right) = \left(\gamma x_1.x_2 + r\right)^d \tag{9}$$

$$K\left(x_1,x_2\right) = e^{-\gamma\left|x_1-x_2\right|^2} \tag{10}$$

$$K\left(x_1,x_2\right) = \tanh\left(\gamma x_1.x_2 + r\right) \tag{11}$$

where x_1 and x_2 are chromaticity data features in vectors form, d is the degree, γ is gamma, and r is the parameter of the kernel projection. Hyperparameters for tuning each model, were iterated, and the model that produced the best accuracy performance were selected and compared.

2.5.3. K-nearest neighbor

KNN algorithm is a non-parametric classifier that uses positional information to categorize or forecast how a single data point will be grouped. The general matric for calculating the distance between data points is Minkowski and for the current application, we used the Euclidean distance formula. The general equation is stated in Equation 12.

$$d\left(i,j\right) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2} \tag{12}$$

where $d\left(i,j\right)$ is the function for calculating the distance between training point i and data point j. x_{i1} and x_{i2} are the chromaticity data of the training set, while x_{j1} and x_{j2} correspond to the chromaticity data of the predictor or validation data.

For model training purpose, the k-value represents the number of closest neighbors and is the primary hyperparameter value for KNN. Since the k-value needed to be established appropriately (46), the value was iterated from 1 to 20, and the k-value with the best performance was discussed.

2.5.4. Decision tree

Decision Tree is a non-parametric supervised learning method for classification and regression to create a model that predicts the value or class of a target variable by learning simple decision rules concluded from the data features. The library provided two criteria settings, "Gini" and "Entropy," to measure the quality of the split in decision rules. The corresponding formulas are stated in Equations (13) and (14).

$$Gini\left(D\right) = 1 - \sum_{i=1}^{k} p_i^2 \tag{13}$$

$$Entrophy\left(D\right) = \sum_{i=1}^{k} - p_i^2 \log_2\left(p_i\right) \tag{14}$$

D corresponds to the dataset, k is the number of classes in the dataset, and p_i is the ratio of the class. Both "Gini" and "Entropy" as provided in the library were utilized for the criterion setting to measure the quality of the split, and the best model was chosen for further analysis and comparison.

2.6. Performance parameter

The model's performance was analyzed using the confusion matrix method based on five parameters: sensitivity, specificity, precision, negative predictive value (NPV), and accuracy (42). The performance of the classification model was evaluated based on the convention stated in the literature. Seven models were trained and validated: Logistic Regression, SVM with Linear, Polynomial, RBF and Sigmoid kernels, KNN, and Decision Tree. The performance of each model was investigated, compared, and analyzed. The implementation of the models in practical applications was also discussed in the present study based on the current application.

3. Results and discussion

This section was organized according to three main subsections; chromaticity analysis, supervised machine learning results, and comparison with other related works. The first phase of analysis revealed the impact of infection on the chromaticity of the chicken comb, and the correlation between chromaticity and health status is discussed. Next, the performance of each developed model is discussed, analyzed, and compared accordingly. Lastly, the performances of all the models are comprehensively compared with reported machine-learning algorithms related to this current application for classifying infected chickens.

3.1. Chromaticity analysis

The difference between healthy and infected chicken comb was illustrated in Figures 4A,B, respectively, using masked images. According to Figure 4, the healthy and infected chicken can be clearly separated based on the chromaticity of the chicken comb, and the impact of infection on the chromaticity value will be further discussed.

The first set of analyses examines the impact of infection on the three-color space parameter and the correlation between each variable parameter. The 3D scatter plot of *x, y,* and *z* data for the pixel-level method and pixel-averaging method are shown in Figures 5A,B.

The scatter plot of the pixel-level method (Figure 5A) appeared to be more complex because of the total data; three sample points from 61 images resulted in 183 points for each class plotted on the graph. However, Figures 5A,B show that both methods have resulted in the same pattern and no significant difference in the distribution of the scatter plot. It can be seen that the infected and healthy chickens were well separated based on the 3D plot. The results were further analyzed by plotting each component in a 2D plot; *xy, xz,* and *yz*. Figures 6A,B present the chromaticity plot of *xy* chromaticity data for pixel-level and pixel-averaging methods, respectively.

Figure 6 shows that the infected and healthy chickens was well separated by x chromaticity for both methods. According to Figures 6A,B, the most infected chicken was scattered below $x = 0.375$,

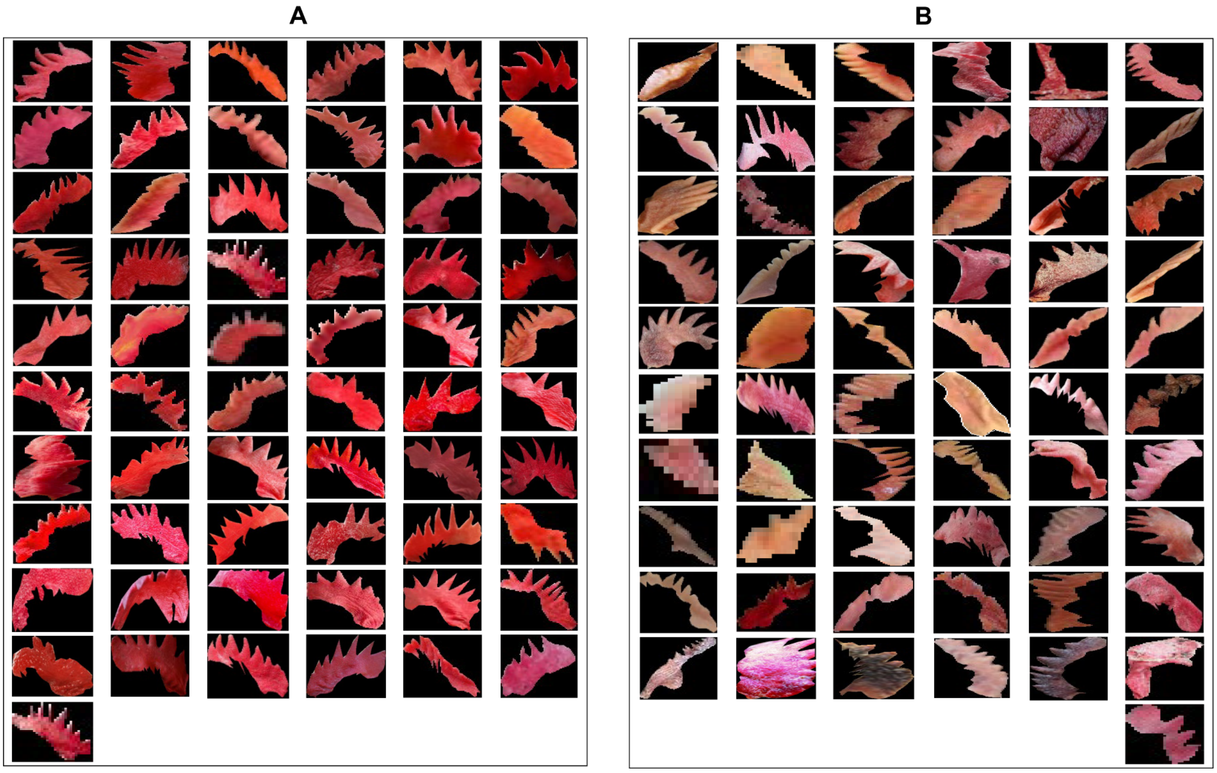


FIGURE 4
(A) Masked images of healthy chicken comb and **(B)** masked images of infected chicken comb.

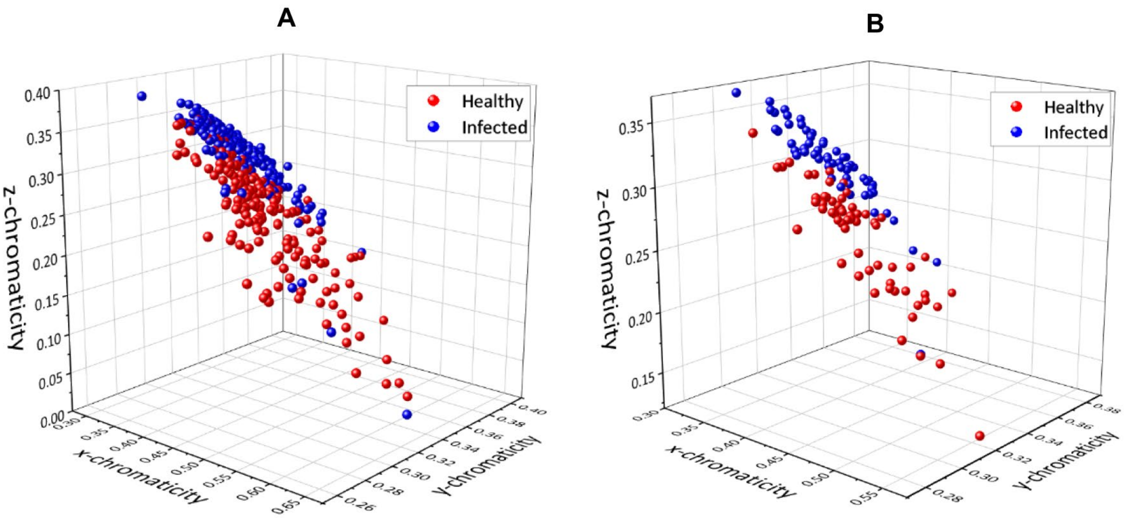


FIGURE 5
3D plots for chromaticity values *x, y,* and *z* using **(A)** pixel-level method and **(B)** pixel-averaging method.

while the healthy chicken was scattered above. The *y* chromaticity value of infected and healthy chickens overlapped, and no specific threshold value can be hypothetically assigned based on the *y* chromaticity variable. However, by combining the *x* and *y* variables, the infected and healthy chickens can be separated more distinctly. Since the scatter plots of healthy and infected chicken were linearly separated, a magenta line was drawn as an indicator line to differentiate between both groups.

Based on the indication line on the pixel-averaging method, it can be observed that only one infected chicken was scattered in the healthy chicken region. On the contrary, 14 infected chickens were spread in the healthy area for the pixel-level method. False classification may occur due to an error in the sampling process. For example, the color of the chicken comb only changes on the front side, and through conventional understanding, the chicken was infected based on that indication. False classification may occur if the sample was taken at

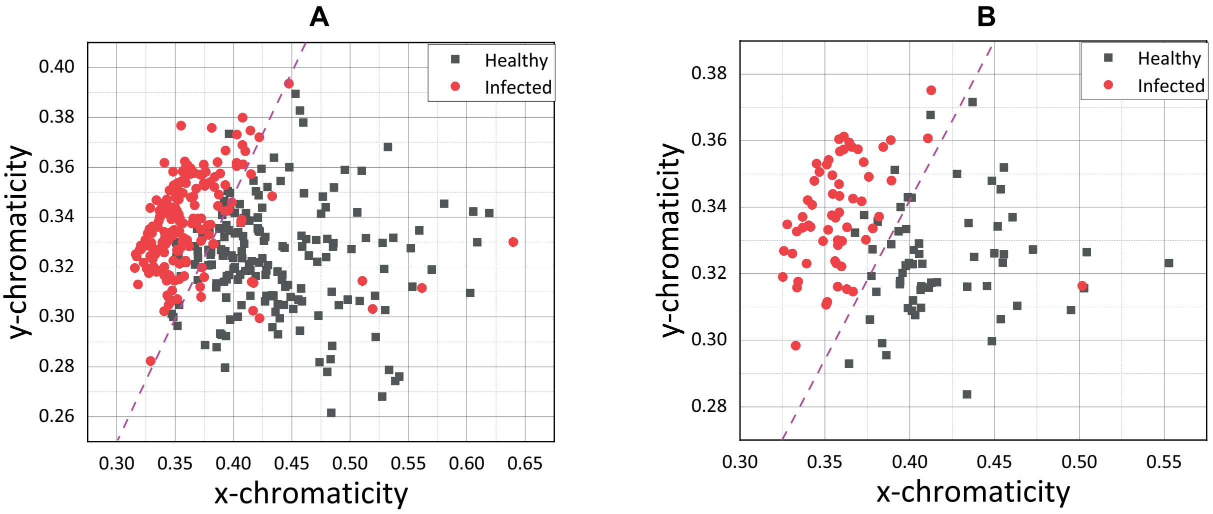


FIGURE 6
(A) *xy* scatter plot for pixel-level method and **(B)** *xy* scatter plot for pixel-averaging method.

the back side of the chicken comb without significant color change. Apart from that, the pixel-averaging method considered all the color data bound in the selected region. Instead of better results in classification, the error and false detection can be reduced. This view was proven by Cao et al. (47), which proposed a new method for water quality detection by considering the average RGB value for the detection (47). Srinivasan et al. (48) also used the average RGB value of each pixel in the image to indicate hemoglobin in human blood for diagnosing anemia.

Since the pixel-averaging method was relevant and gave better results in classifying healthy and infected chickens as shown in Figure 6B, further results and discussion on the impact and correlation between the variables and health status will focus on the pixel-averaging method only. Figures 7A,B show the pixel-averaging methods' results of *xz* and *yz* plots, respectively. Table 1 presents the Logistic Regression's sigmoid input function (referring to Equation 7) according to the pixel-averaging method dataset.

The scatter plot of *xz* (Figure 7A) shows that the infected and healthy chickens can be separated based on the threshold value of below $z = 0.25$ for the *z* chromaticity value. When combining the *x* and *z* chromaticity values, the infected and healthy chicken can be separated based on the magenta line as the hypothetical threshold line. Similarly, by combining *y* and *z* (Figure 7B), the infected and healthy chickens can be classified based on the magenta line drawn. Both plots showed that one chicken could be falsely classified as healthy chicken.

The *x* chromaticity variable was the most dominant variable, followed by z and y variables based on the linear regression sigmoid input function results. It can be observed that the *x* chromaticity variable results in a more significant positive classifier coefficient than the *y* variable with 1.5284 higher by referring to the *xy* model (Table 1). The results show the same trend as in the *xz* model when compared with the *z* chromaticity variable, with 2.2821 higher in the classifier coefficient. Therefore, we can conclude that any small change in the *x* chromaticity variable would significantly contribute to the classification of the chicken. Since the classifier coefficient of *x* chromaticity variable results in a positive sign, the increments of *x* value would increase the value of the sigmoid input function; thus, the results of the sigmoid function would converge to 1. Theoretically, the chroma or actual perceived color was indicated by the *x* and *z* values (43). The *x* chromaticity value can be approximately described as green to red part. So, based on our results in Figures 6B, 7A we conclude that the infected chickens were more converging to green because most of the infected chicken points were scattered below healthy chicken in terms of *x* chromaticity value.

Moving on to the *z* chromaticity variable, the classifier coefficient for the *z* variable was 2.2821 lower when referring to the *xz* model. So, any change in the *z* chromaticity value does not significantly contribute to the classifier predicting the chicken's health. However, according to *yz* model, the *z* chromaticity variable was more dominant than the *y* variable, with 0.6226 higher in the classifier coefficient. Since the coefficient carries a negative sign (*yz* model), increasing the z chromaticity value would encourage the classifier model to predict the chicken to be infected. The *z* chromaticity value can be approximately described as a yellow to blue part for any increment in value (43). Therefore, we conclude that the infected chickens converged more to the blue region according to Figures 7A,B. The weakest variable, *y* chromaticity, has a weaker negative coefficient of 0.8202 compared to the *x* chromaticity variable in the *xy* model. Similarly, in comparison with *z* chromaticity by referring to the *yz* model, the *y* variable resulted in a smaller negative coefficient of 1.5996, while that of the *z* variable was 2.2222. The negative sign indicates that the increased value of y would lower the value of the sigmoid input function; thus, the sigmoid function would converge to 0. The small coefficient of the *y* chromaticity variable was expected based on the *xy* and *yz* plots in Figures 6B, 7B. The scattered point of infected and healthy chicken mostly overlapped in terms of *y* chromaticity value, making the classification nearly impossible. The image data chromaticity's brightness, luminosity, or lightness were represented by the *y* value (43). According to the results, the *y* value was considered the weakest variable that correlated to chicken health due to no significant difference between healthy and infected chickens, and the

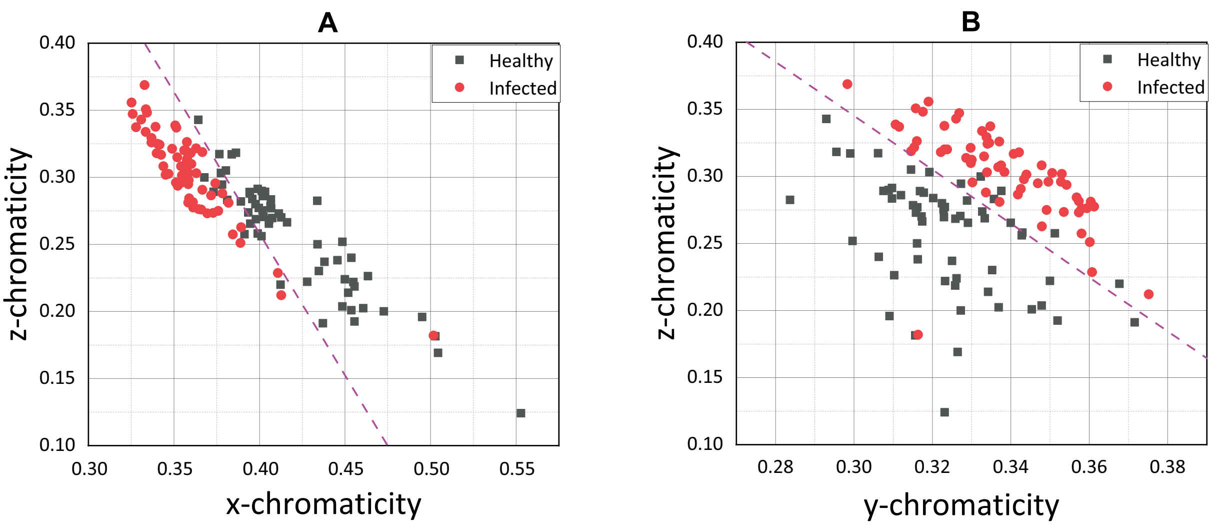


FIGURE 7
(A) *xz* and **(B)** *yz* scatter plot for the pixel-averaging method.

TABLE 1 Sigmoid input function of the Logistic Regression.

Trained variables/Model	Sigmoid input function, $f(x_1,x_2)$
xy	$f(x,y) = 0.29337105 + (2.34857535)x + (-0.8202156)y$
xz	$f(x,z) = 0.29103932 + (2.65420658)x + (0.37211062)z$
yz	$f(y,z) = 0.26176678 + (-1.59957638)y + (-2.22224727)z$

classification was nearly impossible. Therefore, a possible explanation for this might be that our data comes from different sources with different illuminants. This finding corroborated with previous research, which found that the redder comb had more excellent cell-mediated immunity or better health condition (35). Moreover, Martínez-Padilla et al. (49) concluded that comb redness or plasma carotenoids were negatively correlated with *Trichostrongylus tenuis* abundance. Plasma carotenoids are pigments responsible for the vivid color red in the chicken comb, while *T. tenuis* is a nematode in birds that cause diseases.

These findings further support the idea of separating chroma and brightness for the detection method proposed in the literature (47), which uses chromaticity values to measure dissolved water content. However, combining the chromaticity value with the brightness makes the classification viable. The present findings were consistent with previous work (41), which considered intensity and chromaticity features in their algorithm to classify daytime and night images. Furthermore, a study on the correlation between comb color and the immunity system of the chicken was performed based on the red chroma, represented by 600–700 nm, relative to brightness (35).

3.2. Supervised machine learning results

This subsection discusses the performance parameter, advantages, disadvantages, and limitations of all the developed classifiers. Since the Logistic Regression model is the only model that can provide a probability value, the current work will iterate the probability threshold from 0.40 to 0.60, and the expected performance of the model is presented in Figure 8.

The model performance can be categorized into three categories; over-, optimum-, and under-predict the positive event or infected chicken. The first category is over-predicted, which can be seen for the probability threshold of more than 0.53. The model starts to over-predict positive events, resulting in the highest possible sensitivity and NPV of 100% with zero false negative events detected. Secondly, the model can be tuned to get optimum performances which can be indicated by a probability threshold ranging from 0.43 to 0.47. The model was expected to predict 95% for all five performance parameters due to the same amount of false positive and false negative events. Lastly, the proposed model was expected to be under-predicted infected chicken when the probability threshold was below 0.43. The present findings seem consistent with other researchers' views that precision and sensitivity are proportional to actual positive value but have an inverse mutual relationship (50).

Table 2 compares all the supervised machine learning models and notes that the performance of the Logistic Regression was based on an optimum probability threshold of 0.47 and C = 1 for comparison with other models. For SVM models, Linear kernel with C = 1, Polynomial kernel with C = 1 and $d = 1$, RBF kernel with C = 1 and $\gamma = 0.1$, and Sigmoid kernel with C = 1 and $\gamma = 3$ were presented. KNN showed the best performance when the K-value was set more than 5, while for the decision tree model, the Gini criterion was better compared to the Entropy.

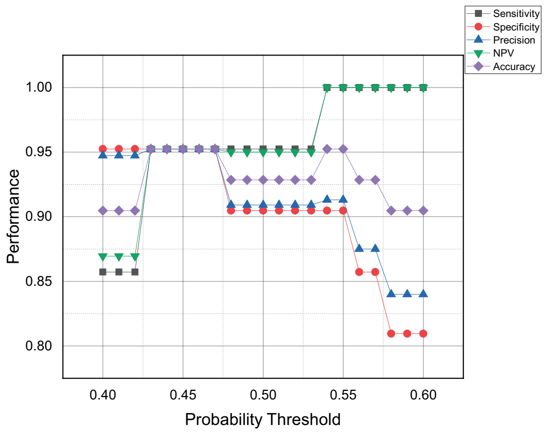


FIGURE 8
The performance parameter of the Logistic Regression model at varying probability threshold.

Logistic Regression, SVM-Linear and Polynomial kernel perform the best in terms of specificity, precision, NPV, and accuracy, followed by SVM-RBF kernel, KNN, Decision Tree, and lastly, SVM-Sigmoidal, in this present study. Logistic Regression, SVM-Linear, and Polynomial kernel perform the best compared to other models because our chromaticity feature data for healthy and infected chicken were linearly separated (Figure 6B). Supporting these statements, researchers in the literature (28) also reported better accuracy using SVM Linear and Polynomial kernel model compared to RBF for their linearly separated dataset.

Incremental learning to extend the model's knowledge while implementing it in practical applications was possible for all models. However, each model has its advantages and disadvantages during implementation. According to the results, the Logistic Regression, SVM Linear and Polynomial kernels perform the best, with a 95% score for all parameters. Compared with other models, logistic regression can output the results in probability values from 0 to 1, and the classification threshold can be assigned. Thus, the performance of the model can be adjusted. However, instead of tunable performance, the stability of the performance itself was an issue during incremental learning (51). Moving on to SVMs models, the storage cost was a significant drawback for these algorithms due to continuous data learning (52, 53). In addition, kernel selection in developing SVM models is essential as it affects the performance of the model. For instance, the SVM-Sigmoid kernel performed at 76% sensitivity, 90% specificity, 88% precision, 79% NPV, and 83% accuracy, which can be considered the lowest among others.

Next, the KNN model has performed similarly to SVM-RBF. KNN model is much simpler than logistic regression and SVM models because it does not need any data training since its algorithms rely on the number of neighbors or K-value for classification. This model's primary limitation is the calculation speed during incremental training (54). False detections may also occur when the data becomes more extensive and no change or update makes for the K-value (55). Another non-parametric model, the decision tree, performed with 86% sensitivity, 95% specificity, 95% precision, and 87% NPV, and 90% accuracy. The Decision Tree is easy to train due to no normalization and data scaling are needed. The algorithms for separating the infected and healthy are intuitive and easy to explain. However, the models may become complex due to the number of depths specified in the training process, and any small change may cause significant changes in the tree's structure (56). Plus, implementing an incremental learning algorithm can variate the stability of the model due to continuous data updates.

In summary, all the models discussed in this subsection can be considered acceptable and successful in classifying health status. Even though current works do not use any specific experimental dataset, all the models have shown to be well developed by just using the randomly well-distributed training and validation image dataset. However, models with high sensitivity, such as Logistic Regression, KNN, SVM-Linear, and SVM-Polynomial, should be considered for current application in providing early warning to prevent major outbreaks. Hicks et al. (50) stated that the consideration of the specificity and precision was based on applications; for medical applications, it is better to over-predict than underestimate the degree of severity. Therefore, current work would consider a model with high sensitivity even though it has a low precision value to over-predict the positive event to prevent significant outbreaks that can cause economic loss and threaten human health.

3.3. Comparison with other reported work

The results reported in this work are compared with other related works which predict the chicken health status. Table 3 shows the summary of the related works. Zhuang et al. (28) utilized an SVM-Polynomial model with 99.469% accuracy to classify infected chicken (bird flu) based on all extracted morphological features: concavity, skeleton altitude angle, shape features, linear area rate, elongation, and circularity. Similarly, other works proposed SVM-RBF models with an accuracy of 97.8% based on extracted locomotor features such as circle variance, elongation, convexity, complexity, eccentricity, and mobility features of walk speed (27). These works (27) were compared with the results reported in this work because they used image processing techniques to extract features as predictors to predict infected chicken. Both works extracted all the morphological, locomotor, and mobility features from the chicken images, and the proposed supervised machine learning classifier model's achieved accuracies >97%. In contrast to these reported works, the results of our work demonstrated that despite only one feature (chicken comb's chromaticity) being used, prediction accuracy as high as 95% can be achieved. This scenario indicates that the chicken comb chromaticity is a very distinctive feature that can be used to predict the bacteria- or virus-infected chickens, as well as confirming the effectiveness of the machine learning models used in this work. It can also be concluded that high prediction accuracy can be achieved with simpler feature extraction and easier image processing technique, if the accurate and distinctive feature is selected.

This reported work is also compared with the deep learning-based algorithms for detecting infected chicken applications. Zhang and Chen (30) have developed a ResNet algorithm with 94% accuracy to detect infected chickens using abnormal swelling images for their training datasets. Other researchers used different textures of chicken-dropping image datasets to classify healthy and infected chickens using XceptionNet with 94% (57) and 98.24% (58) accuracy after fine-tuned. Compared to our works, both of the works (30, 57) have reported lower accuracy. Similar to our work, these works also utilized only one feature, but our reported work utilized a much simpler image

TABLE 2 Comparative analysis of different types of machine learning algorithms.

Model	Confusion matrix	Performance (%)		Model parameters	Data linearity	Incremental learning	Data fitting	Probability output	Performance tuning	Limitation
Logistic Regression	$\begin{bmatrix} 20 & 1 \\ 1 & 20 \end{bmatrix}$	Sensitivity	95	C = 1 Threshold =0.47	Linear	Yes	Yes	Yes	Before and after training	Stability of performance during Incremental training (51)
		Specificity	95							
		Precision	95							
		NPV	95							
		Accuracy	95							
SVM-Linear	$\begin{bmatrix} 20 & 1 \\ 1 & 20 \end{bmatrix}$	Sensitivity	95	C = 1	Linear	Yes	Yes	No	During training	Storage cost from continuous data learning for non-linear SVM (52, 53)
		Specificity	95							
		Precision	95							
		NPV	95							
		Accuracy	95							
SVM-Polynomial	$\begin{bmatrix} 20 & 1 \\ 1 & 20 \end{bmatrix}$	Sensitivity	95	C = 1 $d = 1$	Linear/Non-linear	Yes				
		Specificity	95							
		Precision	95							
		NPV	95							
		Accuracy	95							
SVM-RBF	$\begin{bmatrix} 20 & 1 \\ 2 & 19 \end{bmatrix}$	Sensitivity	95	C = 1 $\gamma = 0.1$	Non-linear	Yes				
		Specificity	90							
		Precision	91							
		NPV	95							
		Accuracy	93							
SVM-Sigmoid	$\begin{bmatrix} 16 & 5 \\ 2 & 19 \end{bmatrix}$	Sensitivity	76	C = 1 $\gamma = 3$	Non-linear	Yes				
		Specificity	90							
		Precision	89							
		NPV	79							
		Accuracy	83							
KNN	$\begin{bmatrix} 20 & 1 \\ 2 & 19 \end{bmatrix}$	Sensitivity	95	k-value = 5	Not applicable	Yes	No	No	Before training	Speed of calculation Data update may deviate (54, 55)
		Specificity	90							
		Precision	91							
		NPV	95							
		Accuracy	93							
Decision tree	$\begin{bmatrix} 18 & 3 \\ 1 & 20 \end{bmatrix}$	Sensitivity	86	Criterion = Gini	Not applicable	Yes	No	No	Before training	Can cause instability for any data change (56)
		Specificity	95							
		Precision	95							
		NPV	87							
		Accuracy	90							

TABLE 3 Summary of related works.

References	Features/Input data	Technique	Model/Algorithms	Performance
Zhuang et al. (28)	Concavity, skeleton altitude angle, shape features, linear area rate, elongation, and circularity	Image processing	SVM Polynomial kernel	99.469% accuracy
Okinda et al. (27)	Circle variance, elongation, convexity, complexity, eccentricity, and walk speed	Image processing	SVM RBF kernel	97.800% accuracy
Zhang and Changxi (30)	Abnormal swelling detection	Deep learning	ResNet	95% accuracy 90% sensitivity
Mbelwa et al. (57)	Abnormal dropping	Deep learning	XceptionNet	94% accuracy
Mbelwa et al. (58)	Abnormal dropping	Deep learning	XceptionNet	98.24% accuracy
Zhuang and Zhang (29)	Chicken images, feather texture, posture	Image processing and deep learning	CNN	99.7% precision

processing technique and lower computational power for training the classifier models. Besides that, Zhuang and Zhang (29) successfully developed algorithms to detect infected chickens with a precision of up to 99.7% by combining image processing and deep learning. To develop these algorithms, authors have utilized the difference in the chicken posture and feather images to train their classifier model. The proposed algorithms were more computationally complex than our work. However, the result performance of the model or classifier was promising. Therefore, it can be proposed that to achieve >99% accuracy, future work will explore on the deep learning algorithms to hybridize with our works to provide early detection algorithms for the prevention of disease outbreaks in poultry farms that can benefit the farmers and improve food safety.

This current work has proven the ability of utilizing the chromaticity of the chicken combs features can be used to detect bacterial- or virus-infected chickens with the help of machine learning models. However, for an implementation in a large-scale chicken farm, a more realistic approach such as capturing the images directly from the chicken cages may be carried out. Further illustration of the accuracy of the model to work in a large -scale poultry farm, by implementing real images dataset, and validation of the model is still needed. Furthermore, hybridization of the chicken comb feature with other established features such as morphological (28), locomotor (27), mobility (27), and optical flow (31), would be future works that need to be considered. The multi-features approach may lead to another breakthrough that would contribute to improved food safety and automation in poultry farm industries.

4. Conclusion

This study presents an early prediction algorithm for detecting bacteria- or virus-infected chickens based on the chromaticity of the chicken comb feature. The algorithm extracted the RGB color data at the area of the chicken comb and converted it into the CIE XYZ color space to analyze the effect of bacteria or virus infection on the chromaticity of the chicken combs. The chromaticity data features of healthy and infected chickens were plotted, and the impact of infection on the chromaticity of the chicken comb was analyzed. Machine learning methods were used to predict the chicken's health status based on the chromaticity feature. The performance analysis of the developed machine learning models proved that the classification of healthy and infected chicken is viable based on the chromaticity of the chicken comb features. All the developed models have excellent generalization to recognize the infected chicken. The results suggest that the chicken comb chromaticity-based algorithm can provide prediction and detection of infected chicken. This algorithm can be applied as a disease monitoring system for the chicken on the farm. In addition, this algorithm can be integrated with other morphological, locomotor, and mobility-based algorithms for detecting infected chickens. Thus, the risk of significant diseases outbreak on the farm could be minimized.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be at: https://github.com/anifakhmal/Infected-vs-Healthy-chick.git.

Author contributions

MAB: methodology, data curation, and writing-original draft preparation. PK: conceptualization, formal analysis, supervision, project administration, writing-review, and editing. ST and AO: supervision, validation, writing-review, and editing. MZB: software, validation, writing-review, and editing. HL: visualization, writing-review, and editing. All authors have read and approved the final manuscript.

Funding

This work was supported by the Ministry of Higher Education, Malaysia, through the Fundamental Research Grant Scheme (FRGS), under the project code of FRGS/1/2020/TK0/UNITEN/02/6 (20210111FRGS). MAB would like to express his gratitude toward the financial support provided by the project under the graduate research assistant scheme. The article processing charge was supported by Tenaga Nasional Berhad (TNB) and UNITEN through the BOLD Refresh Publication Fund under the project code of J510050002-IC-6 BOLDREFRESH2025-Centre of Excellence.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. *Science*. (2010) 327:812–8. doi: 10.1126/science.1185383

2. Liverani M, Waage J, Barnett T, Pfeiffer DU, Rushton J, Rudge JW, et al. Understanding and managing zoonotic risk in the new livestock industries. *Environ Health Perspect*. (2013) 121:873–7. doi: 10.1289/ehp.1206001

3. Jimmy WS. *People, pathogens and our planet volume 1: towards a one health approach for controlling zoonotic diseases* World Bank (2010) Available at: http://hdl.handle.net/10986/2844.

4. Nebehay S. *China's bird flu outbreak cost $6.5 billion* Reuters (2013) Available at: https://www.reuters.com/article/uk-birdflu-idUKBRE94K17U20130521.

5. Greening B, Whitham HK, Aldous WK, Hall N, Garvey A, Mandernach S, et al. Public health response to multistate *Salmonella* Typhimurium outbreak associated with prepackaged chicken salad, United States, 2018. *Int Conf Emerg Infect Dis*. (2022) 28:1254–6. doi: 10.3201/eid2806.211633

6. Gruber P. Avian influenza outbreak approaches largest in US history with 50M birds lost. *Lancaster Farm*. (2022):1–11.

7. Tom P. *Avian flu outbreak wipes out 50.54 mln U.S. birds, a record* Reuters (2022) Available at: https://www.reuters.com/.

8. Lee H, Yoon Y. Etiological agents implicated in foodborne illness world wide. *Food Sci Anim Resour*. (2021) 41:1–7. doi: 10.5851/kosfa.2020.e75

9. Shaw KA, Szablewski CM, Kellner S, Kornegay L, Bair P, Brennan S, et al. Psittacosis outbreak among workers at chicken slaughter plants, Virginia and Georgia, USA, 2018. *Emerg Infect Dis*. (2019) 25:2143–5. doi: 10.3201/eid2511.190703

10. Li J, Fang Y, Qiu X, Yu X, Cheng S, Li N, et al. Human infection with avian-origin H5N6 influenza a virus after exposure to slaughtered poultry. *Emerg Microb Infect*. (2022) 11:807–10. doi: 10.1080/22221751.2022.2048971

11. Hasan AR, Ali M, Siddique M, Rahman M, Islam M. Clinical and laboratory diagnoses of common bacterial diseases of broiler and layer chickens. *Bangladesh J Vet Med*. (2012) 8:107–15. doi: 10.3329/bjvm.v8i2.11188

12. el-Sawah AA, Dahshan ALHM, el-Nahass E-S, el-Mawgoud AIA. Pathogenicity of *Escherichia coli* O157 in commercial broiler chickens. *Beni-Suef Univ J Basic Appl Sci*. (2018) 7:620–5. doi: 10.1016/j.bjbas.2018.07.005

13. Srinivasan P. Bacteriological and pathological studies of egg peritonitis in commercial layer chicken in Namakkal area. *Asian Pacific J Biomed*. (2013) 3:988–94. doi: 10.1016/S2221-1691(13)60191-4

14. Sa-ardta P, Rinder M, Sanyathitiseree P, Weerakhun S, Lertwatcharasarakul P, Lorsunyaluck B, et al. First detection and characterization of Psittaciform bornaviruses in naturally infected and diseased birds in Thailand. *Vet Microbiol*. (2019) 230:62–71. doi: 10.1016/j.vetmic.2019.01.013

15. Zhang Y, Feng B, Xie Z, Deng X, Zhang M, Xie Z, et al. Epidemiological surveillance of parvoviruses in commercial chicken and Turkey farms in Guangxi, southern China, during 2014–2019. *Front Vet Sci*. (2020) 7:561371. doi: 10.3389/fvets.2020.561371

16. Phan LV, Park MJ, Kye SJ, Kim JY, Lee HS, Choi KS. Development and field application of a competitive enzyme-linked immunosorbent assay for detection of Newcastle disease virus antibodies in chickens and ducks. *Poult Sci*. (2013) 92:2034–43. doi: 10.3382/ps.2013-03176

17. Zhao W, Zhu AL, Yu Y, Yuan CL, Zhu CX, Lan DL, et al. Segmentation expression of capsid protein as an antigen for the detection of avian nephritis virus infection in chicken flocks. *J Virol Methods*. (2012) 179:57–61. doi: 10.1016/j.jviromet.2011.09.020

18. Morales Ruiz S, Bendezu J, Choque Guevara R, Montesinos R, Requena D, Choque Moreau L, et al. Development of a lateral flow test for the rapid detection of *Avibacterium paragallinarum* in chickens suspected of having infectious coryza. *BMC Vet Res*. (2018) 14:1–15. doi: 10.1186/s12917-018-1729-0

19. Llarena A-K, Skjerve E, Bjørkøy S, Forseth M, Winge J, Hauge SJ, et al. Rapid detection of *Campylobacter* spp. in chickens before slaughter. *Food Microbiol*. (2022) 103:103949. doi: 10.1016/j.fm.2021.103949

20. Mao Q, Ma S, Schrickel PL, Zhao P, Wang J, Zhang Y, et al. Review detection of Newcastle disease virus. *Front Vet Sci*. (2022) 9:9. doi: 10.3389/fvets.2022.936251

21. Baker D, Jackson EL, Cook S. Perspectives of digital agriculture in diverse types of livestock supply chain systems. Making sense of uses and benefits. *Front Vet Sci*. (2022) 9:992882. doi: 10.3389/fvets.2022.992882

22. Huang F, Xue L, Qi W, Cai G, Liu Y, Lin J. An ultrasensitive impedance biosensor for Salmonella detection based on rotating high gradient magnetic separation and cascade reaction signal amplification. *Biosens Bioelectron*. (2021) 176:176. doi: 10.1016/j.bios.2020.112921

23. Dinshaw IJ, Muniandy S, Teh SJ, Ibrahim F, Leo BF, Thong KL. Development of an aptasensor using reduced graphene oxide chitosan complex to detect *Salmonella*. *J Electroanal Chem*. (2017) 806:88–96. doi: 10.1016/j.jelechem.2017.10.054

24. Cuan K, Zhang T, Huang J, Fang C, Guan Y. Detection of avian influenza-infected chickens based on a chicken sound convolutional neural network. *Comput Electron Agric*. (2020) 178:105688. doi: 10.1016/j.compag.2020.105688

25. Liu L, Li B, Zhao R, Yao W, Shen M, Yang J. A novel method for broiler abnormal sound detection using WMFCC and HMM. *J Sensors*. (2020) 2020:1–7. doi: 10.1155/2020/2985478

26. Carpentier L, Vranken E, Berckmans D, Paeshuyse J, Norton T. Development of sound-based poultry health monitoring tool for automated sneeze detection. *Comput Electron Agric*. (2019) 162:573–81. doi: 10.1016/j.compag.2019.05.013

27. Okinda C, Lu M, Liu L, Nyalala I, Muneri C, Wang J, et al. A machine vision system for early detection and prediction of sick birds: a broiler chicken model. *Biosyst Eng*. (2019) 188:229–42. doi: 10.1016/j.biosystemseng.2019.09.015

28. Zhuang X, Bi M, Guo J, Wu S, Zhang T. Development of an early warning algorithm to detect sick broilers. *Comput Electron Agric*. (2017) 144:102–13. doi: 10.1016/j.compag.2017.11.032

29. Zhuang X, Zhang T. Detection of sick broilers by digital image processing and deep learning. *Biosyst Eng*. (2019) 179:106–16. doi: 10.1016/j.biosystemseng.2019.01.003

30. Zhang H, Changxi Chen. "Design of sick chicken automatic detection system based on improved residual network." in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. Chongqing, China. (2020) 2480–2485.

31. Dawkins MS, Roberts SJ, Cain RJ, Nickson T, Donnelly CA. Early warning of footpad dermatitis and hockburn in broiler chicken flocks using optical flow, bodyweight and water consumption. *Vet Rec*. (2017) 180:499–9. doi: 10.1136/vr.104066

32. Miller PJ. Newcastle disease in poultry. *Aust Vet J*. (1996) 42:138–9. doi: 10.1111/j.1751-0813.1966.tb07645.x

33. Tripathy DN. Fowlpox in chickens and turkeys. *Vet Man*. (2020)

34. Crespo R. Fatty liver hemorrhagic syndrome in poultry. *Vet Man*. (2020) 25:1–3.

35. Mougeot F. Ornamental comb colour predicts T-cell-mediated immunity in male red grouse *Lagopus lagopus* scoticus. *Naturwissenschaften*. (2008) 95:125–32. doi: 10.1007/s00114-007-0303-6

36. Simons MJP, Cohen AA, Verhulst S. What does carotenoid-dependent coloration tell? Plasma carotenoid level signals immunocompetence and oxidative stress state in birds-a meta-analysis. *PLoS One*. (2012) 7:e43088. doi: 10.1371/journal.pone.0043088

37. Guo Z, Ye Q, Li F, Wang Y. Study on corona discharge spatial structure and stages division based on visible digital image colorimetry information. *IEEE Trans Dielectr Electr Insul*. (2019) 26:1448–55. doi: 10.1109/TDEI.2019.008054

38. Jihong L, Fangfan H. "Research and realization of the computer-assisted colorimetry for baked porcelain tooth." in *Proceedings of CIS Work 2007, 2007 International Conference on Computational Intelligence Security Work*. Harbin, China. (2007) 128–131.

39. Wongthanyakram J, Harfield A, Masawat P. A smart device-based digital image colorimetry for immediate and simultaneous determination of curcumin in turmeric. *Comput Electron Agric*. (2019) 166:104981. doi: 10.1016/j.compag.2019.104981

40. Firdaus ML, Saputra E, Ginting SM, Wyantuti S, Eddy DR, Rahmidar L, et al. Smartphone-based digital image colorimetry for non-enzymatic detection of glucose using gold nanoparticles. *Sens Bio Sensing Res*. (2021) 35:100472. doi: 10.1016/j.sbsr.2022.100472

41. Park KH, Lee YS. "Classification of daytime and night based on intensity and chromaticity in RGB color image." in *2018 International Conference on Platform Technology and Service (PlatCon)*. Jeju, Korea (South). (2018) 1–6.

42. Singh P, Singh N, Singh KK, Singh A. Diagnosing of disease using machine learning In: KK Singh, M Elhoseny, A Singh and AA Elngar, editors. *Machine learning and the internet of medical things in healthcare*: Academic Press (2021). 89–111.

43. Poynton CA. "A guided tour of colour space," in *New Foundation for Video Technology: The SMPTE Advanced Television and Electronic Imaging Conference*. San Francisco, CA, USA: IEEE. (1995), 167–1800.

44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30.

45. Belyadi H, Haghighat A. Chapter 5 - supervised learning In: H Belyadi and A Haghighat, editors. *Machine learning guide for oil and gas using Python*: Gulf Professional Publishing (2021). 169–295.

46. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*. (2016) 4:218–8. doi: 10.21037/atm.2016.03.37

47. Cao P, Zhu Y, Zhao W, Liu S, Gao H. Chromaticity measurement based on the image method and its application in water quality detection. *Water*. (2019) 11:2339. doi: 10.3390/w11112339

48. Srinivasan KS, Lakshmi D, Ranganathan H, Gunasekaran N. "Non-invasive estimation of hemoglobin in blood using color analysis." in *1st 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS), 2006*. Tirtayasa, Indonesia. (2006), 547–549.

49. Martínez-Padilla J, Mougeot F, Pérez-Rodríguez L, Bortolotti GR. Nematode parasites reduce carotenoid-based signalling in male red grouse. *Biol Lett*. (2007) 3:161–4. doi: 10.1098/rsbl.2006.0593

50. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. (2022) 12:5979–9. doi: 10.1038/s41598-022-09954-8

51. Paul T, Ueno K. "Robust incremental logistic regression for detection of anomaly using big data." in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA: IEEE. (2020) 1167–1173.

52. Xiao Rong, Wang Jicheng, Zhang Fayan. "An approach to incremental SVM learning algorithm." in *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence ICTAI 2000*. Guangzhou, China: IEEE Computer Society. (2000) 268–273.

53. Kashef R. A boosted SVM classifier trained by incremental learning and decremental unlearning approach. *Expert Syst Appl*. (2021) 167:114154. doi: 10.1016/j.eswa.2020.114154

54. Jie L, Yaxu X, Yadong Y. "Incremental learning algorithm of data complexity based on KNN classifier," in *2020 International Symposium on Community-Centric Systems (CcS)*. Tokyo, Japan: IEEE, (2020) 1–4.

55. Taunk K, De S, Verma S, Swetapadma A. "A brief review of nearest neighbor algorithm for learning and classification." *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. Madurai, India: IEEE. (2019) 1255–1260.

56. Zhang C, Zhang Y, Shi X, Almpanidis G, Fan G, Shen X. On incremental learning for gradient boosting decision trees. *Neural Process Lett*. (2019) 50:957–87. doi: 10.1007/s11063-019-09999-3

57. Mbelwa H, Machuve D, Mbelwa J. Deep convolutional neural network for chicken diseases detection. *Int J Adv Comput Sci Appl*. (2021) 12:759–65. doi: 10.14569/IJACSA.2021.0120295

58. Machuve D, Nwankwo E, Mduma N, Mbelwa J. Poultry diseases diagnostics models using deep learning. *Front Artif Intell*. (2022) 5:5. doi: 10.3389/frai.2022.733345

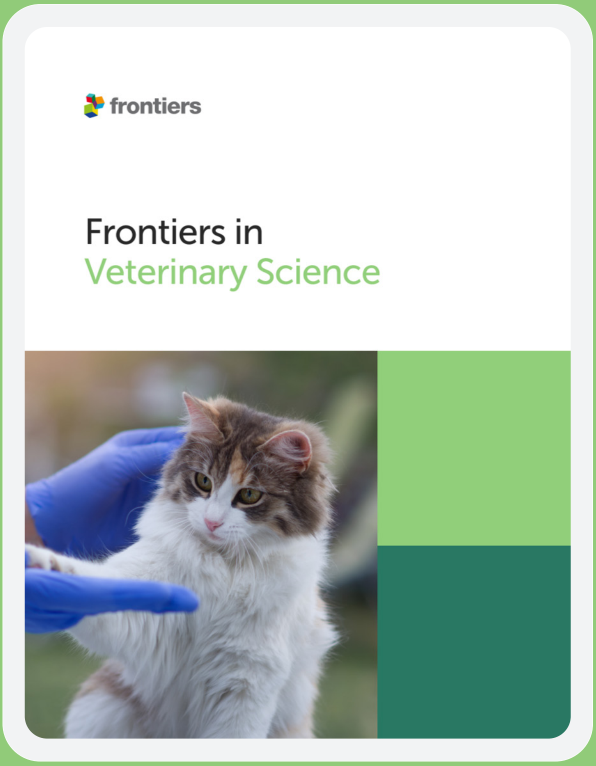
Frontiers in Veterinary Science

Transforms how we investigate and improve animal health

The third most-cited veterinary science journal, bridging animal and human health with a comparative approach to medical challenges. It explores innovative biotechnology and therapy for improved health outcomes.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Veterinary Science

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact